Submitted to the Annals of Statistics

GAUSSIAN APPROXIMATIONS AND MULTIPLIER BOOTSTRAP FOR MAXIMA OF SUMS OF HIGH-DIMENSIONAL RANDOM VECTORS*

By Victor Chernozhukov†, Denis Chetverikov‡ and Kengo Kato§

MIT†, UCLA‡, and University of Tokyo§

We derive a Gaussian approximation result for the maximum of a sum of high dimensional random vectors. Specifically, we establish conditions under which the distribution of the maximum is approximated by that of the maximum of a sum of the Gaussian random vectors with the same covariance matrices as the original vectors. This result applies when the dimension of random vectors (p) is large compared to the sample size (n); in fact, p can be much larger than n, without restricting correlations of the coordinates of these vectors. We also show that the distribution of the maximum of a sum of the random vectors with unknown covariance matrices can be consistently estimated by the distribution of the maximum of a sum of the conditional Gaussian random vectors obtained by multiplying the original vectors with i.i.d. Gaussian multipliers. This is the Gaussian multiplier (or wild) bootstrap procedure. Here too, p can be large or even much larger than n. These distributional approximations, either Gaussian or conditional Gaussian, yield a high-quality approximation to the distribution of the original maximum, often with approximation error decreasing polynomially in the sample size, and hence are of interest in many applications. We demonstrate how our Gaussian approximations and the multiplier bootstrap can be used for modern high dimensional estimation, multiple hypothesis testing, and adaptive specification testing. All these results contain non-asymptotic bounds on approximation errors.

1. Introduction. Let $x_1, \ldots, x_n$ be independent random vectors in $\mathbb{R}^p$, with each x_i having coordinates denoted by x_{ij}, that is, $x_i = (x_{i1}, \ldots, x_{ip})'$. Suppose that each x_i is centered, namely $\mathrm{E}[x_i] = 0$, and has a finite covariance matrix $\mathrm{E}[x_i x_i']$. Consider the rescaled sum:

$$X := (X_1, \ldots, X_p)' := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i. \tag{1}$$

*Date: June, 2012. Revised June, 2013. 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) (25780152), the Japan Society for the Promotion of Science.

AMS 2000 subject classifications: 62E17, 62F40

Keywords and phrases: Dantzig selector, Slepian, Stein method, maximum of vector sums, high dimensionality, anti-concentration

Our goal is to obtain a distributional approximation for the statistic T_0 defined as the maximum coordinate of vector X:

$$T_0 := \max_{1 \leqslant j \leqslant p} X_j.$$

The distribution of T_0 is of interest in many applications. When p is fixed, this distribution can be approximated by the classical Central Limit Theorem (CLT) applied to X. However, in modern applications (cf. [8]), p is often comparable or even larger than n, and the classical CLT does not apply in such cases. This paper provides a tractable approximation to the distribution of T_0 when p can be large and possibly much larger than n.

The *first* main result of the paper is the Gaussian approximation result (GAR), which bounds the Kolmogorov distance between the distributions of T_0 and its Gaussian analog Z_0. Specifically, let $y_1, \dots, y_n$ be independent centered Gaussian random vectors in $\mathbb{R}^p$ such that each y_i has the same covariance matrix as x_i: $y_i \sim N(0, \mathrm{E}[x_i x_i'])$. Consider the rescaled sum of these vectors:

$$Y := (Y_1, \dots, Y_p)' := \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i. \tag{2}$$

Vector Y is the Gaussian analog of X in the sense of sharing the same mean and covariance matrix, namely $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ and $\mathrm{E}[XX'] = \mathrm{E}[YY'] = n^{-1} \sum_{i=1}^n \mathrm{E}[x_i x_i']$. We then define the Gaussian analog Z_0 of T_0 as the maximum coordinate of vector Y:

$$Z_0 := \max_{1 \leqslant j \leqslant p} Y_j. \tag{3}$$

We show that, under suitable moment assumptions, as $n \to \infty$ and possibly $p = p_n \to \infty$,

$$\rho := \sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)| \leqslant C n^{-c} \to 0, \tag{4}$$

where constants $c > 0$ and $C > 0$ are independent of n.

Importantly, in (4), p can be large in comparison to n and be as large as $e^{o(n^c)}$ for some $c > 0$. For example, if x_{ij} are uniformly bounded (namely, $|x_{ij}| \leqslant C_1$ for some constant $C_1 > 0$ for all i and j) the Kolmogorov distance ρ converges to zero at a polynomial rate whenever $(\log p)^7/n \to 0$ at a polynomial rate. We obtain similar results when x_{ij} are sub-exponential and even non-sub-exponential under suitable moment assumptions. Figure 1 illustrates the result (4) in a non-sub-exponential example, which is motivated by the analysis of the Dantzig selector of [9] in non-Gaussian settings (see Section 4).

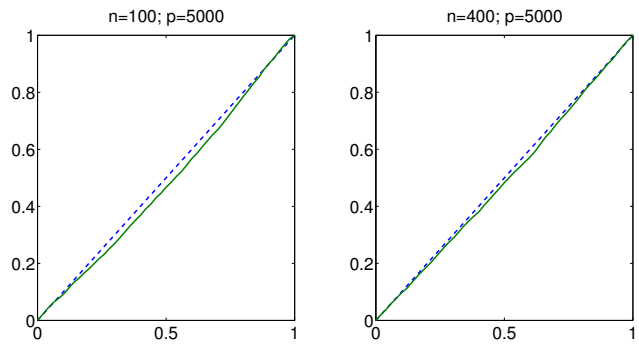

FIG 1. *P-P plots comparing distributions of T_0 and Z_0 in the example motivated by the problem of selecting the penalty level of the Dantzig selector. Here x_{ij} are generated as $x_{ij} = z_{ij}\varepsilon_i$ with $\varepsilon_i \sim t(4)$, (a t-distribution with four degrees of freedom), and z_{ij} are non-stochastic (simulated once using $U[0,1]$ distribution independently across i and j). The dashed line is 45°. The distributions of T_0 and Z_0 are close, as (qualitatively) predicted by the GAR derived in the paper. The quality of the Gaussian approximation is particularly good for the tail probabilities, which is most relevant for practical applications.*

The proof of the Gaussian approximation result (4) builds on a number of technical tools such as Slepian's smart path interpolation (which is related to the solution of Stein's partial differential equation; see Appendix H of the Supplementary Material (SM; [16])), Stein's leave-one-out method, approximation of maxima by the smooth potentials (related to "free energy" in spin glasses) and using some fine or subtle properties of such approximation, and exponential inequalities for self-normalized sums. See, for example, [39, 28, 13, 12, 29, 11, 27, 12, 33] for introduction and prior uses of some of these tools. The proof also critically relies on the anti-concentration and comparison bounds of maxima of Gaussian vectors derived in [11] and restated in this paper as Lemmas 2.1 and 3.1.

Our new Gaussian approximation theorem has the following innovative features. First, we provide a general result that establishes that maxima of sums of random vectors can be approximated in distribution by the maxima of sums of Gaussian random vectors when $p \gg n$ and especially when p is of order $e^{o(n^c)}$ for some $c > 0$. The existing techniques can also lead to results of the form (4) when $p = p_n \to \infty$, but under much stronger conditions on p requiring $p^c/n \to 0$; see Example 17 (Section 10) in [34]. Some high-dimensional cases where p can be of order $e^{o(n^c)}$ can also be handled via Hungarian couplings, extreme value theory or other methods, though special structure is required (for a detailed review, see Section L of the SM [16]). Second, our Gaussian approximation theorem covers cases where T_0 does not have a limit distribution as $n \to \infty$ and $p = p_n \to \infty$. In some cases, after a

suitable normalization, T_0 could have an extreme value distribution as a limit distribution, but the approximation to an extreme value distribution requires some restrictions on the dependency structure among the coordinates in x_i. Our result does not limit the dependency structure. We also emphasize that our theorem specifically covers cases where the process $\{\sum_{i=1}^n x_{ij}/\sqrt{n}, 1 \leqslant j \leqslant p\}$ is not asymptotically Donsker (i.e., can't be embedded into a path of an empirical process that is Donsker). Otherwise, our result would follow from the classical functional central limit theorems for empirical processes, as in [13]. Third, the quality of approximation in (4) is of polynomial order in n, which is better than the logarithmic in n quality that we could obtain in some (though not all) applications using the approximation of the distribution of T_0 by an extreme value distribution (see [31]).

Note that the result (4) 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

$$|\mathrm{P}(T_0 \leqslant c_{Z_0}(1-\alpha)) - (1-\alpha)| \leqslant Cn^{-c} \to 0. \tag{5}$$

The *second* main result of the paper establishes validity of the multiplier (or Wild) 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. Specifically, we define the Gaussian-symmetrized version W_0 of T_0 by multiplying x_i with i.i.d. standard Gaussian random variables $e_1, \ldots, e_n$:

$$W_0 := \max_{1 \leqslant j \leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_i. \tag{6}$$

We show that the conditional quantiles of W_0 given data $(x_i)_{i=1}^n$ are able to 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 obtained in [32]; interesting results on inference on the mean vector of high-dimensional random vectors when $p \gg n$ based

on concentration inequalities and symmetrization are obtained 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.

The key motivating example of our analysis is the analysis of construction of one-sided or two-sided uniform confidence band for high-dimensional means under non-Gaussian assumptions. This requires estimation of a high quantile of the maximum of sample means. We give two concrete applications. One application deals with high-dimensional sparse regression model. In this model, [9] and [6] assume Gaussian errors to analyze the Dantzig selector, where the high-dimensional means enter the constraint in the problem. Our results show that Gaussianity is not necessary and the sharp, 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 among the regressors. This leads to a better choice of the penalty level and tighter bounds on performance than those that had been available previously. In another example we apply our results in the multiple hypothesis testing via the step-down method of [38]. In the SM [16] we also provide an application to adaptive specification testing. In either case the number of hypotheses to be tested or the number of moment restrictions to be tested can be much larger than the sample size. Lastly, in a companion work ([10]), we derive 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 [30, 7, 19, 26] for the entire empirical process and have applications to inference in nonparametric problems such as construction of uniform confidence bands and testing qualitative hypotheses (see, e.g., [25], [21], and [18]).

1.1. *Organization of the paper.* In Section 2, we give the results on Gaussian approximation, and in Section 3 on the multiplier bootstrap. In Sections 4 and 5, we develop applications to the Dantzig selector and multiple testing. Appendices A-C contain proofs for each of these sections, with Appendix A stating auxiliary tools and lemmas. Due to the space limitation, we put additional results and proofs into the SM [16]. In particular, Appendix M of the SM provides additional application to adaptive specification testing. Results of Monte Carlo simulations are presented in Appendix G of the SM.

1.2. *Notation.* In what follows, unless otherwise stated, we will assume that $p \geqslant 3$. In making asymptotic statements, we assume that $n \to \infty$ with understanding that p depends on n and possibly $p \to \infty$ as $n \to \infty$. Constants $c, C, c_1, C_1, c_2, C_2, \dots$ are understood to be independent of

n. Throughout the paper, $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leqslant i \leqslant n$, that is, 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 $z \in \mathbb{R}^p$, z' denotes the transpose of z. 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 $x = (x_1, \ldots, x_p)'$. We 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 $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$. For two sets A and B, $A \ominus B$ denotes their symmetric difference, that is, $A \ominus B = (A\backslash B) \cup (B\backslash A)$.

2. Gaussian Approximations for Maxima of Non-Gaussian Sums. The purpose of this section is to compare and bound the difference between the expectations and distribution functions of the non-Gaussian to Gaussian maxima:

$$T_0 := \max_{1\leqslant j\leqslant p} X_j \quad \text{and} \quad Z_0 := \max_{1\leqslant j\leqslant p} Y_j,$$

where vector X is defined in equation (1) and Y in equation (2). Here and in what follows, without loss of generality, we will assume that $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent. In order to derive the main result of this section, we shall employ Slepian interpolation, Stein's leave-one-out method, a truncation method combined with self-normalization, as well as some fine properties of the smooth max function (such as "stability"). (The relative complexity of the approach is justified in Comment 2.5 below.)

The following bounds on moments will be used in stating the bounds in Gaussian approximations:

$$M_k := \max_{1\leqslant j\leqslant p} (\bar{\mathrm{E}}[|x_{ij}|^k])^{1/k}. \tag{7}$$

The problem of comparing distributions of maxima is of intrinsic difficulty since the maximum function $z = (z_1, \ldots, z_p)' \mapsto \max_{1\leqslant j\leqslant p} z_j$ is nondifferentiable. 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:

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

where $\beta > 0$ is the smoothing parameter that controls the level of approximation (we call this function the "smooth max function"). An elementary

calculation shows that for all $z \in \mathbb{R}^p$,

$$0 \leqslant F_\beta(z) - \max_{1 \leqslant j \leqslant p} z_j \leqslant \beta^{-1} \log p. \tag{8}$$

This smooth max function arises in the definition of "free energy" in spin glasses; see, for example, [29]. Some important properties of this function, such as stability, are derived in the Appendix.

Given a threshold level $u > 0$, we define a truncated version of x_{ij} by

$$\tilde{x}_{ij} = x_{ij} 1\left\{|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\right\} - \mathrm{E}\left[x_{ij} 1\left\{|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\right\}\right]. \tag{9}$$

Let $\varphi_x(u)$ be the infimum, which is attained, over all numbers $\varphi \geqslant 0$ such that

$$\bar{\mathrm{E}}\left[x_{ij}^2 1\left\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\right\}\right] \leqslant \varphi^2 \bar{\mathrm{E}}[x_{ij}^2]. \tag{10}$$

Note that the function $\varphi_x(u)$ is right-continuous; it measures the impact of truncation on second moments. Define $u_x(\gamma)$ as the infimum over all numbers $u \geqslant 0$ such that

$$\mathrm{P}\left(|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\right) \geqslant 1 - \gamma.$$

Also define $\varphi_y(u)$ and $u_y(\gamma)$ by the corresponding quantities for the analogue Gaussian case, namely with $(x_i)_{i=1}^n$ replaced by $(y_i)_{i=1}^n$ in the above definitions. Throughout the paper we use the following quantities:

$$\varphi(u) := \varphi_x(u) \vee \varphi_y(u), \quad u(\gamma) := u_x(\gamma) \vee u_y(\gamma).$$

Also, in what follows, for a smooth function $g : \mathbb{R} \to \mathbb{R}$, write

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

The following theorem is the main building block toward deriving a result of the form (4).

THEOREM 2.1 (Comparison of Gaussian to Non-Gaussian Maxima). *Let $\beta > 0, u > 0$ and $\gamma \in (0,1)$ be such that $2\sqrt{2} u M_2 \beta / \sqrt{n} \leqslant 1$ and $u \geqslant u(\gamma)$. Then for every $g \in C_b^3(\mathbb{R})$, $|\mathrm{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim D_n(g, \beta, u, \gamma)$, so that*

$$|\mathrm{E}[g(T_0) - g(Z_0)]| \lesssim D_n(g, \beta, u, \gamma) + \beta^{-1} G_1 \log p,$$

where

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

We will also invoke the following lemma, which is proved in [11].

LEMMA 2.1 (Anti-Concentration). *(a) Let $Y_1, \dots, Y_p$ be jointly Gaussian random variables with $\mathrm{E}[Y_j] = 0$ and $\sigma_j^2 := \mathrm{E}[Y_j^2] > 0$ for all $1 \leqslant j \leqslant p$, and let $a_p := \mathrm{E}[\max_{1 \leqslant j \leqslant p}(Y_j/\sigma_j)]$. 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 $\varsigma > 0$,*

$$\sup_{z \in \mathbb{R}} \mathrm{P}\left(|\max_{1 \leqslant j \leqslant p} Y_j - z| \leqslant \varsigma \right) \leqslant C\varsigma\{a_p + \sqrt{1 \vee \log(\underline{\sigma}/\varsigma)}\},$$

where $C > 0$ is a constant depending only on $\underline{\sigma}$ and $\bar{\sigma}$. When σ_j are all equal, $\log(\underline{\sigma}/\varsigma)$ on the right side can be replaced by 1. (b) Furthermore, the worst case bound is obtained by bounding a_p by $\sqrt{2 \log p}$.

By Theorem 2.1 and Lemma 2.1, we can obtain a bound on the Kolmogorov distance, ρ, between the distribution functions of T_0 and Z_0, which is the main theorem of this section.

THEOREM 2.2 (**Main Result 1: Gaussian Approximation**). *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 $\gamma \in (0, 1)$,*

$$\rho \leqslant C \left\{ n^{-1/8}(M_3^{3/4} \vee M_4^{1/2})(\log(pn/\gamma))^{7/8} + n^{-1/2}(\log(pn/\gamma))^{3/2} u(\gamma) + \gamma \right\},$$

where $C > 0$ is a constant that depends on c_1 and C_1 only.

COMMENT 2.1 (Removing lower bounds on the variance). The condition that $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$ for *all* $1 \leqslant j \leqslant p$ can not be removed in general. However, this condition becomes redundant, if there is at least a nontrivial fraction of components x_{ij}'s of vector x_i with variance bounded away from zero and all pairwise correlations bounded away from 1: for some $J \subset \{1, \dots, p\}$,

$$|J| \geqslant \nu p, \quad \bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1, \quad \frac{|\bar{\mathrm{E}}[x_{ij} x_{ik}]|}{\sqrt{\bar{\mathrm{E}}[x_{ij}^2]}\sqrt{\bar{\mathrm{E}}[x_{ik}^2]}} \leqslant 1 - \nu', \quad \forall (k, j) \in J \times J : k \neq j,$$

where $\nu > 0$ and $\nu' > 0$ are some constants independent of n or p. Section J of the SM [16] contains formal results under this condition. ■

In applications, it is useful to have explicit bounds on the upper function $u(\gamma)$. To this end, let $h : [0, \infty) \to [0, \infty)$ be a *Young-Orlicz modulus*, that is, a convex and strictly increasing function with $h(0) = 0$. Denote by h^{-1} the inverse function of h. Standard examples include the power function $h(v) = v^q$ with inverse $h^{-1}(\gamma) = \gamma^{1/q}$ and the exponential function $h(v) = \exp(v) -$

1 with inverse $h^{-1}(\gamma) = \log(\gamma + 1)$. These functions describe how many moments the random variables have; for example, a random variable ξ has finite qth moment if $\mathrm{E}[|\xi|^q] < \infty$, and is sub-exponential if $\mathrm{E}[\exp(|\xi|/C)] < \infty$ for some $C > 0$. We refer to [30], Chapter 2.2, for further details.

LEMMA 2.2 (Bounds on the upper function $u(\gamma)$). *Let $h : [0, \infty) \to [0, \infty)$ be a Young-Orlicz modulus, and let $B > 0$ and $D > 0$ be constants such that $(\mathrm{E}[x_{ij}^2])^{1/2} \leqslant B$ for all $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p$, and $\bar{\mathrm{E}}[h(\max_{1\leqslant j\leqslant p} |x_{ij}|/D)] \leqslant 1$. Then under the condition of Theorem 2.2,*

$$u(\gamma) \leqslant C \max\{Dh^{-1}(n/\gamma), B\sqrt{\log(pn/\gamma)}\},$$

where $C > 0$ is a constant that depends on c_1 and C_1 only.

In applications, parameters B and D (with M_3 and M_4 as well) are allowed to increase with n. The size of these parameters and the choice of the Young-Orlicz modulus are case-specific.

2.1. *Examples.* The purpose of this subsection is to obtain bounds on ρ for various leading examples frequently encountered in applications. We are concerned with simple conditions under which ρ decays polynomially in n.

Let $c_1 > 0$ and $C_1 > 0$ be some constants, and let $B_n \geqslant 1$ be a sequence of constants. We allow for the case where $B_n \to \infty$ as $n \to \infty$. We shall first consider applications where one of the following conditions is satisfied *uniformly in* $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$:

$$\text{(E.1)}\quad c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1 \text{ and } \max_{k=1,2} \bar{\mathrm{E}}[|x_{ij}|^{2+k}/B_n^k] + \mathrm{E}[\exp(|x_{ij}|/B_n)] \leqslant 4;$$

$$\text{(E.2)}\quad c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1 \text{ and } \max_{k=1,2} \bar{\mathrm{E}}[|x_{ij}|^{2+k}/B_n^k] + \mathrm{E}[(\max_{1\leqslant j\leqslant p} |x_{ij}|/B_n)^4] \leqslant 4.$$

COMMENT 2.2. As a rather special case, Condition (E.1) covers vectors x_i made up from sub-exponential random variables, that is,

$$\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1 \text{ and } \mathrm{E}[\exp(|x_{ij}|/C_1)] \leqslant 2$$

(set $B_n = C_1$), which in turn includes, as a special case, vectors x_i made up from sub-Gaussian random variables. Condition (E.1) also covers the case when $|x_{ij}| \leqslant B_n$ for all i and j, where B_n may increase with n. Condition (E.2) is weaker than (E.1) in that it restricts only the growth of the fourth moments but stronger than (E.1) in that it restricts the growth of $\max_{1\leqslant j\leqslant p} |x_{ij}|$. ∎

We shall also consider regression applications where one of the following conditions is satisfied *uniformly in* $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$:

(E.3) $x_{ij} = z_{ij}\varepsilon_{ij}$, where z_{ij} are non-stochastic with $|z_{ij}| \leqslant B_n$, $\mathbb{E}_n[z_{ij}^2] = 1$, and $\mathrm{E}[\varepsilon_{ij}] = 0$, $\mathrm{E}[\varepsilon_{ij}^2] \geqslant c_1$, and $\mathrm{E}[\exp(|\varepsilon_{ij}|/C_1)] \leqslant 2$; or

(E.4) $x_{ij} = z_{ij}\varepsilon_{ij}$, where z_{ij} are non-stochastic with $|z_{ij}| \leqslant B_n$, $\mathbb{E}_n[z_{ij}^2] = 1$, and $\mathrm{E}[\varepsilon_{ij}] = 0$, $\mathrm{E}[\varepsilon_{ij}^2] \geqslant c_1$, and $\mathrm{E}[\max_{1\leqslant j\leqslant p} \varepsilon_{ij}^4] \leqslant C_1$.

Comment 2.3. Conditions (E.3) and (E.4) cover examples that arise in high-dimensional regression, for example, [9], which we shall revisit later in the paper. Typically, ε_{ij}'s are independent of j (i.e., $\varepsilon_{ij} = \varepsilon_i$) and hence $\mathrm{E}[\max_{1\leqslant j\leqslant p} \varepsilon_{ij}^4] \leqslant C_1$ in condition (E.4) reduces to $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$. Interestingly, these examples are also connected to spin glasses, see, for example, [29] and [33] (z_{ij} can be interpreted as generalized products of "spins" and ε_i as their random "interactions"). Note that conditions (E.3) and (E.4) are special cases of conditions (E.1) and (E.2) but we state (E.3) and (E.4) explicitly because these conditions are useful in applications. ∎

Corollary 2.1 (**Gaussian Approximation in Leading Examples**). *Suppose that there exist constants $c_2 > 0$ and $C_2 > 0$ such that one of the following conditions is satisfied: (i) (E.1) or (E.3) holds and $B_n^2(\log(pn))^7/n \leqslant C_2 n^{-c_2}$ or (ii) (E.2) or (E.4) holds and $B_n^4(\log(pn))^7/n \leqslant C_2 n^{-c_2}$. Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, C_1, c_2, and C_2 such that*

$$\rho \leqslant Cn^{-c}.$$

Comment 2.4. This corollary follows relatively directly from Theorem 2.2 with help of Lemma 2.2. Moreover, from Lemma 2.2, it is routine to find other conditions that lead to the conclusion of Corollary 2.1. ∎

Comment 2.5 (The benefits from the overall proof strategy). We note in Section I of the SM [16], that it is possible to derive the following result by a much simpler proof:

Lemma 2.3 (A Simple GAR). *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 there exists a constant $C > 0$ depending only on c_1 and C_1 such that*

$$\sup_{t\in\mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)| \leqslant C(n^{-1}(\log(pn))^7)^{1/8}(\bar{\mathrm{E}}[S_i^3])^{1/4}, \tag{11}$$

where $S_i := \max_{1\leqslant j\leqslant p}(|x_{ij}| + |y_{ij}|)$.

This simple (though apparently new, at this level of generality) result follows from the classical Lindeberg's argument previously given in Chatterjee

[6] (in the special context of a spin-glass setting like (E.4) with $\epsilon_{ij} = \epsilon_i$) in combination with Lemma 2.1 and standard kernel smoothing of indicator functions. In the SM [16], we provide the proof using Slepian-Stein methods, which a reader wishing to see a simple exposition (before reading a much more involved proof of the main results) may find helpful. The bound here is only useful in some limited cases, for example, in (E.3) or (E.4) when $B_n^6(\log(pn))^7/n \to 0$. When $B_n^6(\log(pn))^7/n \to \infty$, the simple methods fail, requiring a more delicate argument. Note that in applications B_n typically grows at a fractional power of n, see, for example, [10] and [17], and so the limitation is rather major, and was the principal motivation for our whole paper. ∎

3. Gaussian Multiplier Bootstrap.

3.1. *A Gaussian-to-Gaussian Comparison Lemma.* The proofs of the main results in this section rely on the following lemma. Let V and Y be centered Gaussian random vectors in $\mathbb{R}^p$ with covariance matrices Σ^V and Σ^Y, respectively. The following lemma compares the distribution functions of $\max_{1\leqslant j\leqslant p} V_j$ and $\max_{1\leqslant j\leqslant p} Y_j$ in terms of p and

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

LEMMA 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 $1 \leqslant j \leqslant p$. Then there exists a constant $C > 0$ depending only on c_1 and C_1 such that*

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

COMMENT 3.1. The result is derived in [11], and extends that of [11] who gave an explicit error in Sudakov-Fernique comparison of expectations of maxima of Gaussian random vectors. ∎

3.2. *Results on Gaussian Multiplier Bootstrap.* Suppose that we have a dataset $(x_i)_{i=1}^n$ consisting of n independent centered random vectors x_i in $\mathbb{R}^p$. In this section, we are interested in approximating quantiles of

$$T_0 = \max_{1\leqslant j\leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} \tag{12}$$

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

$$W_0 = \max_{1 \leqslant j \leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_i. \tag{13}$$

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

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

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.

Before presenting the theorem, we first give a simple useful lemma that is helpful in the proof of the theorem and in power analysis in applications. 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 $(y_i)_{i=1}^n$ is a sequence of independent $N(0, \mathrm{E}[x_i x_i'])$ vectors. Recall that $\Delta = \max_{1 \leqslant j,k \leqslant p} \big|\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]\big|$.

Lemma 3.2 (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 $\alpha \in (0,1)$,*

$$\mathrm{P}\big(c_{W_0}(\alpha) \leqslant c_{Z_0}(\alpha + \pi(\vartheta))\big) \geqslant 1 - \mathrm{P}(\Delta > \vartheta),$$
$$\mathrm{P}\big(c_{Z_0}(\alpha) \leqslant c_{W_0}(\alpha + \pi(\vartheta))\big) \geqslant 1 - \mathrm{P}(\Delta > \vartheta),$$

where, for $C_2 > 0$ denoting a constant depending only on c_1 and C_1,

$$\pi(\vartheta) := C_2 \vartheta^{1/3} (1 \vee \log(p/\vartheta))^{2/3}.$$

Recall that $\rho := \sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)|$. We are now in position to state the first main theorem of this section.

Theorem 3.1 (**Main Result 2: Validity of Multiplier Bootstrap for High-Dimensional Means**). *Suppose that for some constants $0 < c_1 < C_1$, we have $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Then for every $\vartheta > 0$,*

$$\rho_\ominus := \sup_{\alpha \in (0,1)} \mathrm{P}(\{T_0 \leqslant c_{W_0}(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \leqslant 2(\rho + \pi(\vartheta) + \mathrm{P}(\Delta > \vartheta)),$$

where $\pi(\cdot)$ is defined in Lemma 3.2. In addition,

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) - \alpha| \leqslant \rho_\ominus + \rho.$$

Theorem 3.1 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 $T = T(x_1 \dots, x_n)$ which may not be of the form (12), can be approximated by T_0 of the form (12), and that the multiplier bootstrap is performed on a statistic $W = W(x_1, \dots, x_n, e_1, \dots, e_n)$, which may be different from (13) but still can be approximated by W_0 of the form (13).

We require the approximation to hold in 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

$$\mathrm{P}(|T - T_0| > \zeta_1) < \zeta_2, \tag{14}$$

$$\mathrm{P}(\mathrm{P}_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2. \tag{15}$$

We use the α-quantile of $W = W(x_1, \dots, x_n, e_1, \dots, e_n)$, 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.

Lemma 3.3 (Comparison of Quantiles, II). *Suppose that condition (15) is satisfied. Then for every $\alpha \in (0,1)$,*

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

The next result provides a bound on the bootstrap estimation error.

Theorem 3.2 (**Main Result 3: Validity of Multiplier Bootstrap for Approximate High-Dimensional Means**). *Suppose that, for some constants $0 < c_1 < C_1$, we have $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Moreover, suppose that (14) and (15) hold. Then for every $\vartheta > 0$,*

$$\begin{aligned}\rho_\ominus &:= \sup_{\alpha \in (0,1)} \mathrm{P}(\{T \leqslant c_W(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \\ &\leqslant 2(\rho + \pi(\vartheta) + \mathrm{P}(\Delta > \vartheta)) + C_3 \zeta_1 \sqrt{1 \vee \log(p/\zeta_1)} + 5\zeta_2,\end{aligned}$$

where $\pi(\cdot)$ is defined in Lemma 3.2, and $C_3 > 0$ depends only on c_1 and C_1. In addition, $\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant \rho_\ominus + \rho$.

COMMENT 3.2 (On Empirical and other bootstraps). In this paper, we focus on the Gaussian multiplier bootstrap (which is a form of wild bootstrap). This is because other exchangeable bootstrap methods are asymptotically equivalent to this bootstrap. For example, consider the empirical (or Efron's) bootstrap which approximates the distribution of T_0 by the conditional distribution of $T_0^* = \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n (x_{ij}^* - \mathbb{E}_n[x_{ij}])/\sqrt{n}$ where $x_1, \ldots, x_n^*$ are i.i.d. draws from the empirical distribution of $x_1, \ldots, x_n$. We show in Section K of the SM [16], that the empirical bootstrap is asymptotically equivalent to the Gaussian multiplier bootstrap, by virtue of Theorem 2.2 (applied conditionally on the data). The validity of the empirical bootstrap then follows from the validity of the Gaussian multiplier method. The result is demonstrated under a simplified condition. A detailed analysis of more sophisticated conditions, and the validity of more general exchangeably weighted bootstraps (see [35]) in the current setting, will be pursued in future work. ■

3.3. *Examples Revisited.* Here we revisit the examples in Section 2.1 and see how the multiplier bootstrap works for these leading examples. Let, as before, $c_2 > 0$ and $C_2 > 0$ be some constants, and let $B_n \geqslant 1$ be a sequence of constants. Recall conditions (E.1)-(E.4) in Section 2.1. The next corollary shows that the multiplier bootstrap is valid with a polynomial rate of accuracy for the significance level under weak conditions.

COROLLARY 3.1 (**Multiplier Bootstrap in Leading Examples**). *Suppose that conditions (14) and (15) hold with $\zeta_1 \sqrt{\log p} + \zeta_2 \leqslant C_2 n^{-c_2}$. Moreover, suppose that one of the following conditions is satisfied: (i) (E.1) or (E.3) holds and $B_n^2 (\log(pn))^7/n \leqslant C_2 n^{-c_2}$ or (ii) (E.2) or (E.4) holds and $B_n^4 (\log(pn))^7/n \leqslant C_2 n^{-c_2}$. Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, C_1, c_2, and C_2 such that*

$$\rho_\ominus = \sup_{\alpha \in (0,1)} \mathrm{P}(\{T \leqslant c_W(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \leqslant C n^{-c}.$$

In addition, $\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant \rho_\ominus + \rho \leqslant C n^{-c}$.

4. Application: Dantzig Selector in the Non-Gaussian Model. The purpose of this section is to demonstrate the case with which the GAR and the multiplier bootstrap theorem given in Corollaries 2.1 and 3.1 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 [9], [6], [43] in the Gaussian setting and of [29] in a sub-exponential 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 consider both homoscedastic and heteroscedastic models.

4.1. *Homoscedastic case.* Let $(z_i, y_i)_{i=1}^n$ be a sample of independent observations where $z_i \in \mathbb{R}^p$ is a non-stochastic 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 in such a way that $\mathbb{E}_n[z_{ij}^2] = 1$. Here we consider the homoscedastic case:

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

where σ^2 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: $\|\cdot\|_I$, where the label I is the name of a norm of interest. 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 jth component seminorm for δ is $\|\delta\|_{\mathrm{jc}} := |\delta_j|$, and so on.

The Dantzig selector is the estimator defined by

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \|b\|_{\ell_1} \text{ subject to } \sqrt{n} \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'b)]| \leqslant \lambda, \tag{16}$$

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 confidence level $1 - \alpha$ (where α is a number close to zero.) 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 Gaussian approximations 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 [5], 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 [9] and [6]. 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 GAR 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 $\tilde{z}_i = (z_i', -z_i')'$ where z_i' denotes the transpose of z_i.

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 4.1 (A comment on the identifiability factor $\kappa_I(\beta)$). The identifiability factors $\kappa_I(\beta)$ depend on the true parameter value β. These factors represent a modest generalization of the cone invertibility factors and sensitivity characteristics defined in [43] and [24], which are known to be quite general. The difference 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

$$\kappa_{\mathrm{pr}}(\beta) \geqslant 2^{-1} s^{-1/2} \kappa(s,1), \tag{17}$$

where $\kappa(s,1)$ is the restricted eigenvalue defined in [6]. ■

Next we state 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 4.1 (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 $B_n \geqslant 1$ of constants such that for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$: (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) $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_1}$. 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 only four bounded moments. However, the probabilistic guarantee is not $1-\alpha$ as, for example, in [6], but rather $1 - \alpha - Cn^{-c}$, which reflects the cost of non-Gaussianity (along with more stringent side conditions). In what follows we discuss details of this result. Note that the bound above holds for any semi-norm of interest $\|\cdot\|_I$.

COMMENT 4.2 (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 $\sqrt{n}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}e_i]| = O_{\mathrm{P}}(1)$, so that $c_{Z_0}(1-\alpha) = O(1)$, whereas $c_0(1-\alpha) \propto \sqrt{\log p}$. Thus, the performance guarantee provided by $\widehat{\beta}^{(1)}$ can be much better than that of $\widehat{\beta}^{(0)}$. ∎

Comment 4.3 (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 $\lambda = c_{Z_0}(1-\alpha)$. Suppose that the true value β is sparse, namely $\|\beta\|_0 \leqslant s$. In this case, with probability at least $1-\alpha-Cn^{-c}$,

$$\|\widehat{\beta}^{(1)} - \beta\|_{\mathrm{pr}} \leqslant \frac{2c_{Z_0}(1-\alpha)}{\sqrt{n}\kappa_{\mathrm{pr}}(\beta)} \leqslant \frac{4\sqrt{s}c_0(1-\alpha)}{\sqrt{n}\kappa(s,1)} \leqslant \frac{4\sqrt{s}\sqrt{2\log(\alpha/(2p))}}{\sqrt{n}\kappa(s,1)}, \tag{18}$$

where the last bound is the same as in [6], 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 [6], but rather $1-\alpha-Cn^{-c}$, which together with side conditions is the cost of non-Gaussianity. ∎

Comment 4.4 (Other refinements). Unrelated to the main theme of this paper, we can see from (18) 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. ∎

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

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

that is, 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 GAR applies as before, namely the difference of the distribution functions of T_0 and its Gaussian analogue Z_0 converges to zero. In this case, the Gaussian analogue 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 is the upper bound on the error variance assumed to be known. Let $(e_i)_{i=1}^n$ be a sequence of i.i.d. standard Gaussian 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 (16) with $\lambda = \lambda^{(1)} = c_W(1-\alpha)$ to obtain $\widehat{\beta}^{(1)}$.

THEOREM 4.2 (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 $B_n \geqslant 1$ of constants such that for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$: (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) $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_1}$; and (vi) $(\log p)B_n c_0(1-1/n)/(\sqrt{n}\kappa_{\mathrm{pr}}(\beta)) \leqslant C_1 n^{-c_1}$. 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*

$$\|\widehat{\beta}^{(1)} - \beta\|_I \leqslant \frac{2\lambda^{(1)}}{\sqrt{n}\kappa_I(\beta)}. \tag{19}$$

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; where $c_{Z_0}(1-a) \leqslant c_0(1-a)$.

COMMENT 4.5 (A Portmanteu Signicance Test). The result above contains a practical test of joint significance of all regressors, that is, a test of the hypothesis that $\beta_0 = 0$, with the exact asymptotic size α.

COROLLARY 4.1. *Under conditions of the either of preceding two theorems, the test, that rejects the null hypothesis $\beta_0 = 0$ if $\widehat{\beta}^{(1)} \neq 0$, has size equal to $\alpha + Cn^{-c}$.*

To see this note that under the null hypothesis of $\beta_0 = 0$, β_0 satisfies the constraint in (16) with probability $(1 - \alpha - Cn^{-c})$, by construction of λ; hence $\|\widehat{\beta}^{(1)}\| \leqslant \|\beta_0\| = 0$ with exactly this probability. Appendix M of the SM [16] generalizes this to a more general test, which tests $\beta_0 = 0$ in the regression model $y_i = d_i'\gamma_0 + x_i'\beta_0 + \varepsilon_i$, where d_i's are a small set of variables, whose coefficients are not known and need to be estimated. The test orthogonalizes each x_{ij} with respect to d_i by partialling out linearly the effect of d_i on x_{ij}. The result similar to that in the corollary continues to hold. ∎

Comment 4.6 (Confidence Bands). Following Gautier and Tsybakov [24], the bounds given in the preceding theorems can be used for Scheffe-type (simultaneous) inference on all components of β_0.

Corollary 4.2. *Under the conditions of either of the two preceding theorems, a $(1-\alpha-Cn^{-c})$-confidence rectangle for β_0 is given by the region $\times_{j=1}^p I_j$, where $I_j = [\widehat{\beta}_j^{(1)} \pm 2\lambda^{(1)}/(\sqrt{n}\kappa_{jc}(\beta)]$.*

We note that $\kappa_{\mathrm{jc}}(\beta) = 1$ if $\mathbb{E}_n[z_{ij}z_{ik}] = 0$ for all $k \neq j$. Therefore, in the orthogonal model of Donoho and Johnstone, where $\mathbb{E}_n[z_{ij}z_{ik}] = 0$ for all pairs $j \neq k$, we have that $\kappa_{\mathrm{jc}}(\beta) = 1$ for all $1 \leqslant j \leqslant p$, so that $I_j = [\widehat{\beta}_j^{(1)} \pm 2\lambda^{(1)}/\sqrt{n}]$, which gives a practical simultaneous $(1 - \alpha - Cn^{-c})$ confidence rectangle for β. In non-orthogonal designs, we can rely on [24]'s tractable linear programming algorithms for computing lower bounds on $\kappa_I(\beta)$ for various norms I of interest; see also [27]. ∎

Comment 4.7 (Generalization of Dantzig Selector). There are many interesting applications where the results given above apply. There are, for example, interesting works by [1] and [23] 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 (see references cited above for both of these points). 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 $\mathcal{R}(\beta) = \{0\}$. ■

5. Application: Multiple Hypothesis Testing via the Stepdown Method. In this section, we study the problem of multiple hypothesis testing in the framework of multiple means or, more generally, approximate means. The latter possibility allows us to cover the case of testing multiple coefficients in multiple regressions, which is often required in empirical studies; see, for example, [2]. We combine a general stepdown procedure described in [38] with the multiplier bootstrap developed in this paper. In contrast with [38], our results do not require weak convergence arguments, and, thus, can be applied to models with an increasing number of means. Notably, the number of means can be large in comparison with the sample size.

Let $\beta := (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ be a vector of parameters of interest. We are interested in simultaneously testing the set of null hypotheses $H_j : \beta_j \leqslant \beta_{0j}$ against the alternatives $H_j' : \beta_j > \beta_{0j}$ for $j = 1, \ldots, p$ where $\beta_0 := (\beta_{01}, \ldots, \beta_{0p})' \in \mathbb{R}^p$. Suppose that the estimator $\widehat{\beta} := (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)' \in \mathbb{R}^p$ is available that has an approximately linear form:

$$\sqrt{n}(\widehat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i + r_n, \tag{20}$$

where $x_1, \ldots, x_n$ are independent zero-mean random vectors in $\mathbb{R}^p$, the influence functions, and $r_n := (r_{n1}, \ldots, r_{np})' \in \mathbb{R}^p$ are linearization errors that are small in the sense required by condition (M) below. Vectors $x_1, \ldots, x_n$ need not be directly observable. Instead, some estimators $\widehat{x}_1, \ldots, \widehat{x}_n$ of influence functions $x_1, \ldots, x_n$ are available, which will be used in the bootstrap simulations.

We refer to this framework as testing multiple approximate means. This framework covers the case of testing multiple means with $r_n = 0$. More generally, this framework also covers the case of multiple linear and nonlinear m-regressions; see, for example, [26] for explicit conditions giving rise to linearizaton (20). The detailed exposition of how the case of multiple linear regressions fits into this framework can be found in [13]. Note also that this framework implicitly covers the case of testing equalities ($H_j : \beta_j = \beta_{0j}$) because equalities can be rewritten as pairs of inequalities.

We are interested in a procedure 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 $\alpha + o(1)$ uniformly over a large class of data-generating processes and, in particular, 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_j is equivalent to $\omega \in \Omega_j$ for some subset Ω_j of Ω. Let $\mathcal{W} := \{1, \dots, p\}$ and for $w \subset \mathcal{W}$ denote $\Omega^w := (\cap_{j \in w} \Omega_j) \cap (\cap_{j \notin w} \Omega_j^c)$ where $\Omega_j^c := \Omega \backslash \Omega_j$. The strong control of the family-wise error rate means

(21)
$$\sup_{w \subset \mathcal{W}} \sup_{\omega \in \Omega^w} \mathrm{P}_\omega\{\text{reject at least one hypothesis among } H_j,\ j \in w\} \leqslant \alpha + o(1)$$

where P_ω denotes the probability distribution under the data-generating process ω. This setting is clearly of interest in many empirical studies.

For $j = 1, \dots, p$, denote $t_j := \sqrt{n}(\widehat{\beta}_j - \beta_{0j})$. The stepdown procedure of [38] 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 $\max_{j \in w} t_j$. On the first step, let $w(1) = \mathcal{W}$. Reject all hypotheses H_j satisfying $t_j > c_{1-\alpha,w(1)}$. If no null hypothesis is rejected, then stop. If some H_j are rejected, 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_j, $j \in w(l)$, satisfying $t_j > c_{1-\alpha,w(l)}$. If no null hypothesis is rejected, then stop. If some H_j are rejected, let $w(l+1)$ be the subset of all null hypotheses among $j \in w(l)$ that were not rejected. Proceed in this way until the algorithm stops.

Romano and Wolf [38] proved the following result. Suppose that $c_{1-\alpha,w}$ satisfy

$$c_{1-\alpha,w'} \leqslant c_{1-\alpha,w''} \quad \text{whenever } w' \subset w'', \tag{22}$$

$$\sup_{w \subset \mathcal{W}} \sup_{\omega \in \Omega^w} \mathrm{P}_\omega \left(\max_{j \in w} t_j > c_{1-\alpha,w} \right) \leqslant \alpha + o(1), \tag{23}$$

then inequality (21) holds if the stepdown procedure is used. 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_{j_0} be this hypothesis. Clearly, we have $w(l) \supset w$. So,

$$\max_{j \in w} t_j \geqslant t_{j_0} > c_{1-\alpha,w(l)} \geqslant c_{1-\alpha,w}.$$

Combining this chain of inequalities with (23) yields (21).

To obtain suitable $c_{1-\alpha,w}$ that satisfy inequalities (22) and (23) 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 $\sum_{i=1}^n \widehat{x}_{ij} e_i / \sqrt{n}$ given $(\widehat{x}_i)_{i=1}^n$.

To prove that so defined critical values $c_{1-\alpha,w}$ satisfy inequalities (22) and (23), the following two quantities play a key role:

$$\Delta_1 := \max_{1\leqslant j\leqslant p} |r_{nj}| \text{ and } \Delta_2 := \max_{1\leqslant j\leqslant p} \mathbb{E}_n[(\widehat{x}_{ij} - x_{ij})^2].$$

We will assume the following regularity condition,

(M) There are positive constants c_2 and C_2: (i) $\mathrm{P}\left(\sqrt{\log p}\Delta_1 > C_2 n^{-c_2}\right) < C_2 n^{-c_2}$ and (ii) $\mathrm{P}\left((\log(pn))^2\Delta_2 > C_2 n^{-c_2}\right) < C_2 n^{-c_2}$. In addition, one of the following conditions is satisfied: (iii) (E.1) or (E.3) holds and $B_n^2(\log(pn))^7/n \leqslant C_2 n^{-c_2}$ or (iv) (E.2) or (E.4) holds and $B_n^4(\log(pn))^7/n \leqslant C_2 n^{-c_2}$.

THEOREM 5.1 (Strong Control of Family-Wise Error Rate). *Suppose that (M) is satisfied uniformly over a class of data-generating processes Ω. Then the stepdown procedure with the multiplier bootstrap critical values $c_{1-\alpha,w}$ given above satisfy (21) for this Ω with $o(1)$ strengthened to Cn^{-c} for some constants $c > 0$ and $C > 0$ depending only on c_1, C_1, c_2, and C_2.*

COMMENT 5.1 (The case of sample means). Let us consider the simple case of testing multiple means. In this case, $\beta_j = \mathrm{E}[z_{ij}]$ and $\widehat{\beta}_j = \mathbb{E}_n[z_{ij}]$, where $z_i = (z_{ij})_{j=1}^p$ are i.i.d. vectors, so that the influence functions are $x_{ij} = z_{ij} - \mathrm{E}[z_{ij}]$, and the remainder is zero, $r_n = 0$. The influence functions x_i are not directly observable, though easily estimable by demeaning, $\widehat{x}_{ij} = z_{ij} - \mathbb{E}_n[z_{ij}]$ for all i and j. It is instructive to see the implications of Theorem 5.1 in this simple setting. Condition (i) of assumption (M) holds trivially in this case. Condition (ii) of assumption (M) follows from Lemma A.1 under conditions (iii) or (iv) of assumption (M). Therefore, Theorem 5.1 applies provided that $\underline{\sigma}^2 \leqslant \mathrm{E}[x_{ij}^2] \leqslant \bar{\sigma}^2$, $(\log p)^7 \leqslant C_2 n^{1-c_2}$ for arbitrarily small c_2 and, for example, either (a) $\mathrm{E}[\exp(|x_{ij}|/C_1)] \leqslant 2$ (condition (E.1)) or (b) $\mathrm{E}[\max_{1\leqslant j\leqslant p} x_{ij}^4] \leqslant C_1$ (condition (E.2)). Hence, the theorem implies that the Gaussian multiplier bootstrap as described above leads to a testing procedure with the strong control of the family-wise error rate for the multiple hypothesis testing problem of which the *logarithm* of the number of hypotheses is nearly of order $n^{1/7}$. Note here that no assumption that limits the dependence between $x_{i1}, \ldots, x_{ip}$ or the distribution of x_i is made. Previously, [4] proved strong control of the family-wise error rate for the Rademacher multiplier bootstrap with some adjustment factors assuming that x_i's are Gaussian with unknown covariance structure. ■

COMMENT 5.2 (Relation to Simultaneous Testing). The question on how large p can be was studied in [22] 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-statistics t_j and assuming that t_j were distributed as, say, $N(0,1)$. This framework is conservative in that correlation amongst statistics is dealt away by independence, namely by Šidák procedures. In contrast, our approach takes into account the correlation amongst statistics and hence is asymptotically exact, that is, asymptotically non-conservative. ■

APPENDIX A: PRELIMINARIES

A.1. A Useful Maximal Inequality. The following lemma, which is derived in [11], is a useful variation of standard maximal inequalities.

LEMMA A.1 (Maximal Inequality). *Let $x_1, \ldots, x_n$ be independent random vectors in $\mathbb{R}^p$ with $p \geqslant 2$. Let $M = \max_{1 \leqslant i \leqslant n} \max_{1 \leqslant j \leqslant p} |x_{ij}|$ 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. See [11], Lemma 8. ■

A.2. Properties of the Smooth Max Function. We will use the following properties of the smooth max function.

LEMMA A.2 (Properties of F_β). *For every $1 \leqslant j, k, l \leqslant p$,*

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

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

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

Moreover,

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

PROOF OF LEMMA A.2. The first property was noted in [11]. The other properties follow from repeated application of the chain rule. ■

LEMMA A.3 (Lipschitz Property of F_β). *For every $x \in \mathbb{R}^p$ 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.3. The proof follows from the fact that $\partial_j F_\beta(z) = \pi_j(z)$ with $\pi_j(z) \geqslant 0$ and $\sum_{j=1}^p \pi_j(z) = 1$. ∎

We will also use the following properties of $m = g \circ F_\beta$. We assume $g \in C_b^3(\mathbb{R})$ in Lemmas A.4-A.6 below.

Lemma A.4 (Three derivatives of $m = g \circ F_\beta$). *For every $1 \leqslant j, k, l \leqslant p$,*

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

where π_j, w_{jk} and q_{jkl} are defined in Lemma A.2, and (z) denotes evaluation at z, including evaluation of F_β at z.

Proof of lemma A.4. The proof follows from repeated application of the chain rule and by the properties noted in Lemma A.2. ∎

Lemma A.5 (Bounds on derivatives of $m = g \circ F_\beta$). *For every $1 \leqslant j, k, l \leqslant p$,*

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

where

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

Moreover,

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

Proof of Lemma A.5. The lemma follows from a direct calculation. ∎

The following lemma plays a critical role.

Lemma A.6 (Stability Properties of Bounds over Large Regions). *For every $z \in \mathbb{R}^p$, $w \in \mathbb{R}^p$ with $\max_{j\leqslant p}|w_j|\beta \leqslant 1$, $\tau \in [0,1]$, and every $1 \leqslant j, k, l \leqslant p$, we have*

$$U_{jk}(z) \lesssim U_{jk}(z + \tau w) \lesssim U_{jk}(z),\ U_{jkl}(z) \lesssim U_{jkl}(z + \tau w) \lesssim U_{jkl}(z).$$

Proof of Lemma A.6. Observe that

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

Similarly, $\pi_j(z+\tau w) \geqslant e^{-2}\pi_j(z)$. Since U_{jk} and U_{jkl} are finite sums of products of terms such as π_j, π_k, π_l, δ_{jk}, the claim of the lemma follows. ∎

A.3. Lemma on Truncation. The proof of Theorem 2.1 uses the following properties of the truncation operation. Define $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$ and $\tilde{X} = n^{-1/2}\sum_{i=1}^n \tilde{x}_i$, where "tilde" denotes the truncation operation defined in Section 2. The following lemma also covers the special case where $(x_i)_{i=1}^n = (y_i)_{i=1}^n$. The property (d) is a consequence of sub-Gaussian inequality of [12], Theorem 2.16, for self-normalized sums.

Lemma A.7 (Truncation Impact). *For every $1 \leqslant j,k \leqslant p$ and $q \geqslant 1$, (a) $(\bar{\mathrm{E}}[|\tilde{x}_{ij}|^q])^{1/q} \leqslant 2(\bar{\mathrm{E}}[|x_{ij}|^q])^{1/q}$; (b) $\bar{\mathrm{E}}[|\tilde{x}_{ij}\tilde{x}_{ik} - x_{ij}x_{ik}|] \leqslant (3/2)(\bar{\mathrm{E}}[x_{ij}^2] + \bar{\mathrm{E}}[x_{ik}^2])\varphi(u)$; (c) $\mathbb{E}_n[(\mathrm{E}[x_{ij}1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\}])^2] \leqslant \bar{\mathrm{E}}[x_{ij}^2]\varphi^2(u)$. Moreover, for a given $\gamma \in (0,1)$, let $u \geqslant u(\gamma)$ where $u(\gamma)$ is defined in Section 2. Then: (d) with probability at least $1-5\gamma$, for all $1 \leqslant j \leqslant p$,*

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

Proof. See Section D of SM [16]. ∎

APPENDIX B: PROOFS FOR SECTION 2

B.1. Proof of Theorem 2.1. The second claim of the theorem follows from property (8) of the smooth max function. Hence we shall prove the first claim. The proof strategy is similar to the proof of Lemma I.1. However, to control effectively the third order terms in the leave-one-out expansions we shall use truncation and 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 let

$$\tilde{y}_i := (\tilde{y}_{ij})_{j=1}^p,\ \tilde{y}_{ij} := y_{ij}1\left\{|y_{ij}| \leqslant u(\bar{\mathrm{E}}[y_{ij}^2])^{1/2}\right\},\ \tilde{Y} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{y}_i.$$

Note that by the symmetry of the distribution of y_{ij}, $\mathrm{E}[\tilde{y}_{ij}] = 0$. Recall that we are assuming 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 Appendix A.

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

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

where $\Delta(\gamma, u) := 5M_2\varphi(u)\sqrt{2\log(p/\gamma)}$. By Lemma A.7, we have $\mathrm{E}[\mathcal{I}] \geqslant 1 - 10\gamma$. Observe that by Lemma A.3,

$$|m(x) - m(y)| \leqslant G_1|F_\beta(x) - F_\beta(y)| \leqslant G_1 \max_{1\leqslant j\leqslant p}|x_j - y_j|,$$

so that

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

hence

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

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

$$|\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)M_3^3 + (G_2 + \beta G_1)M_2^2\varphi(u).$$

We define the Slepian interpolation $Z(t)$ between $\tilde{Y}$ and $\tilde{Z}$, Stein's leave-one-out version $Z^{(i)}(t)$ of $Z(t)$, and other useful terms:

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

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

We have by Taylor's theorem,

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

where

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

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

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

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 $I = 0$. Moreover, in Steps 3 and 4 below, we will show that

$$|II| \lesssim (G_2+\beta G_1)M_2^2\varphi(u),\ |III| \lesssim n^{-1/2}(G_3+G_2\beta+G_1\beta^2)M_3^3.$$

The claim of this step now follows.

Step 3. (Bound on II) By independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)Z_{ik}(t)$,

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

where the last step follows from Lemma A.5. Since $|\sqrt{t}\tilde{x}_{ij}+\sqrt{1-t}\tilde{y}_{ij}| \leqslant 2\sqrt{2}uM_2$, so that $|\beta(\sqrt{t}\tilde{x}_{ij}+\sqrt{1-t}\tilde{y}_{ij})/\sqrt{n}| \leqslant 1$ (which is satisfied by the assumption $\beta 2\sqrt{2}uM_2/\sqrt{n} \leqslant 1$), by Lemmas A.6 and A.5, the last expression is bounded up to an absolute constant by

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

Observe that since $\mathrm{E}[x_{ij}x_{ik}] = \mathrm{E}[y_{ij}y_{ik}]$, we have that $\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)] = n^{-1}\mathrm{E}[\tilde{x}_{ij}\tilde{x}_{ik}-\tilde{y}_{ij}\tilde{y}_{ik}] = n^{-1}\mathrm{E}[\tilde{x}_{ij}\tilde{x}_{ik}-x_{ij}x_{ik}]+n^{-1}\mathrm{E}[y_{ij}y_{ik}-\tilde{y}_{ij}\tilde{y}_{ik}]$, so that

by Lemma A.7 (b), $\sum_{i=1}^n |\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]| \leqslant \bar{\mathrm{E}}[|\tilde{x}_{ij}\tilde{x}_{ik} - x_{ij}x_{ik}|] + \bar{\mathrm{E}}[|y_{ij}y_{ik} - \tilde{y}_{ij}\tilde{y}_{ik}|] \lesssim (\bar{\mathrm{E}}[x_{ij}^2] + \bar{\mathrm{E}}[x_{ik}^2])\varphi(u) \lesssim M_2^2\varphi(u)$. Therefore, we conclude that $|II| \lesssim (G_2 + G_1\beta)M_2^2\varphi(u)$.

Step 4. (Bound on III) Observe that

$$
\begin{aligned}
|III| &\leqslant_{(1)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t) + \tau Z_i(t))|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]d\tau dt \\
&\lesssim_{(2)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t))|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&=_{(3)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t))] \cdot \mathrm{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt,
\end{aligned} \tag{24}
$$

where (1) follows from $|\partial_j\partial_k\partial_l m(z)| \leqslant U_{jkl}(z)$ (see Lemma A.5), (2) from Lemma A.6, (3) from independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)$. Moreover, the last expression is bounded as follows:

$$
\begin{aligned}
&\text{right-hand side of (24)} \lesssim_{(4)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \mathrm{E}[U_{jkl}(Z(t))] \cdot \mathrm{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&=_{(5)} \sum_{j,k,l=1}^p \int_0^1 \mathrm{E}[U_{jkl}(Z(t))] \cdot n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&\leqslant_{(6)} \int_0^1 \left(\sum_{j,k,l=1}^p \mathrm{E}[U_{jkl}(Z(t))] \right) \max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&\lesssim_{(7)} (G_3 + G_2\beta + G_1\beta^2) \int_0^1 \max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt,
\end{aligned}
$$

where (4) follows from Lemma A.6, (5) from definition of $\bar{\mathrm{E}}$, (6) from a trivial inequality, (7) from Lemma A.5. We have to bound the integral on the last line. Let $\omega(t) = 1/(\sqrt{t} \wedge \sqrt{1-t})$, and observe that

$$
\begin{aligned}
&\int_0^1 \max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&= \int_0^1 \omega(t) \max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|(\dot{Z}_{ij}(t)/\omega(t))Z_{ik}(t)Z_{il}(t)|]dt \\
&\leqslant n\int_0^1 \omega(t) \max_{1\leqslant j,k,l\leqslant p} \left(\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)/\omega(t)|^3]\bar{\mathrm{E}}[|Z_{ik}(t)|^3]\bar{\mathrm{E}}[|Z_{il}(t)|^3]\right)^{1/3} dt,
\end{aligned}
$$

where the last inequality is by Hölder. The last term is further bounded as

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

where (1) follows from the fact that: $|\dot{Z}_{ij}(t)/\omega(t)| \leqslant (|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)/\sqrt{n}$, $|Z_{im}(t)| \leqslant (|\tilde{x}_{im}|+|\tilde{y}_{im}|)/\sqrt{n}$, and the product of terms $\bar{\mathrm{E}}[(|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)^3]^{1/3}$, $\bar{\mathrm{E}}[(|\tilde{x}_{ik}|+|\tilde{y}_{ik}|)^3]^{1/3}$ and $\bar{\mathrm{E}}[(|\tilde{x}_{il}|+|\tilde{y}_{il}|)^3]^{1/3}$ is trivially bounded by $\max_{1\leqslant j\leqslant p}$ $\bar{\mathrm{E}}[(|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)^3]$; (2) follows from $\int_0^1\omega(t)dt \lesssim 1$, (3) from Lemma A.7 (a), and (4) from the normality of y_{ij} with $\mathrm{E}[y_{ij}^2]=\mathrm{E}[x_{ij}^2]$, so that $\mathrm{E}[|y_{ij}|^3] \lesssim (\mathrm{E}[y_{ij}^2])^{3/2} = (\mathrm{E}[|x_{ij}^2|])^{3/2} \leqslant \mathrm{E}[|x_{ij}|^3]$. This completes the overall proof. ■

B.2. Proof of Theorem 2.2. See Appendix D.2 of the SM [16]. ■

B.3. Proof of Lemma 2.2. Since $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$ by assumption, we have $1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\} \leqslant 1\{|x_{ij}| > c_1^{1/2}u\}$. By Markov's inequality and the condition of the lemma, we have

$$\begin{aligned}
&\mathrm{P}\left(|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2},\ \text{for some } (i,j)\right) \leqslant \textstyle\sum_{i=1}^n \mathrm{P}\left(\max_{1\leqslant j\leqslant p}|x_{ij}| > c_1^{1/2}u\right)\\
&\quad\leqslant \textstyle\sum_{i=1}^n \mathrm{P}\left(h(\max_{1\leqslant j\leqslant p}|x_{ij}|/D) > h(c_1^{1/2}u/D)\right) \leqslant n/h(c_1^{1/2}u/D).
\end{aligned}$$

This implies $u_x(\gamma) \leqslant c_1^{-1/2}Dh^{-1}(n/\gamma)$. For $u_y(\gamma)$, by $y_{ij}\sim N(0,\mathrm{E}[x_{ij}^2])$ with $\mathrm{E}[x_{ij}^2]\leqslant B^2$, we have $\mathrm{E}[\exp(y_{ij}^2/(4B^2))] \lesssim 1$. Hence

$$\begin{aligned}
&\mathrm{P}\left(|y_{ij}| > u(\bar{\mathrm{E}}[y_{ij}^2])^{1/2},\ \text{for some } (i,j)\right) \leqslant \textstyle\sum_{i=1}^n\sum_{j=1}^p \mathrm{P}(|y_{ij}| > c_1^{1/2}u)\\
&\quad\leqslant \textstyle\sum_{i=1}^n\sum_{j=1}^p \mathrm{P}(|y_{ij}|/(2B) > c_1^{1/2}u/(2B)) \lesssim np\exp(-c_1u^2/(4B^2)).
\end{aligned}$$

Therefore, $u_y(\gamma) \leqslant CB\sqrt{\log(pn/\gamma)}$ where $C>0$ depends only on c_1. ■

B.4. Proof of Corollary 2.1. Since conditions (E.3) and (E.4) are special cases of (E.1) and (E.2), it suffices to prove the result under conditions (E.1) and (E.2) only. The proof consists of two steps.

Step 1. In this step, in each case of conditions (E.1) and (E.2), we shall compute the following bounds on moments M_3 and M_4 and parameters B and D in Lemma 2.2 with specific choice of h:

(E.1) $B \vee M_3^3 \vee M_4^2 \leqslant CB_n$, $D \leqslant CB_n \log p$, $h(v) = e^v - 1$;
(E.2) $B \vee D \vee M_3^3 \vee M_4^2 \leqslant CB_n$, $h(v) = v^4$;

Here $C > 0$ is a (sufficiently large) constant that depends only on c_1 and C_1. The bounds on B, M_3 and M_4 follow from elementary computations using Hölder's inequality. The bounds on D follow from an elementary application of Lemma 2.2.2 in [30]. For brevity, we omit the detail.

Step 2. In all cases, there are sufficiently small constants $c_3 > 0$ and $c_4 > 0$, and a sufficiently large constant $C_3 > 0$, depending only on c_1, C_1, c_2, C_2 such that, with $\ell_n := \log(pn^{1+c_3})$,

$$n^{-1/2}\ell_n^{3/2} \max\{B\ell_n^{1/2}, Dh^{-1}(n^{1+c_3})\} \leqslant C_3 n^{-c_4},$$
$$n^{-1/8}(M_3^{3/4} \vee M_4^{1/2})\ell_n^{7/8} \leqslant C_3 n^{-c_4}.$$

Hence taking $\gamma = n^{-c_3}$, we conclude from Theorem 2.2 and Lemma 2.2 that $\rho \leqslant Cn^{-\min\{c_3,c_4\}}$ where $C > 0$ depends only on c_1, C_1, c_2, C_2. ∎

APPENDIX C: PROOFS FOR SECTION 3

C.1. Proof of Lemma 3.2. 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 Lemma 3.1, on the event $\{(x_i)_{i=1}^n : \Delta \leqslant \vartheta\}$, we have $|\mathrm{P}(Z_0 \leqslant t) - \mathrm{P}_e(W_0 \leqslant t)| \leqslant \pi(\vartheta)$ for all $t \in \mathbb{R}$, and so on this event

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

implying the first claim. The second claim follows similarly. ∎

C.2. Proof of Lemma 3.3. By equation (15), 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. ∎

C.3. Proof of Theorem 3.1. For $\vartheta > 0$, let $\pi(\vartheta) := C_2\vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3}$ as defined in Lemma 3.2. To prove the first inequality, note that

$$\begin{aligned}
&\mathrm{P}(\{T_0 \leqslant c_{W_0}(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \\
&\leqslant_{(1)} \mathrm{P}(c_{Z_0}(\alpha - \pi(\vartheta)) < T_0 \leqslant c_{Z_0}(\alpha + \pi(\vartheta))) + 2\mathrm{P}(\Delta > \vartheta) \\
&\leqslant_{(2)} \mathrm{P}(c_{Z_0}(\alpha - \pi(\vartheta)) < Z_0 \leqslant c_{Z_0}(\alpha + \pi(\vartheta))) + 2\mathrm{P}(\Delta > \vartheta) + 2\rho \\
&\leqslant_{(3)} 2\pi(\vartheta) + 2\mathrm{P}(\Delta > \vartheta) + 2\rho,
\end{aligned}$$

where (1) follows from Lemma 3.2, (2) follows from the definition of ρ, and (3) follows from the fact that Z_0 has no point masses. The first inequality follows. The second inequality follows from the first inequality and the definition of ρ. ∎

C.4. Proof of Theorem 3.2. For $\vartheta > 0$, let $\pi(\vartheta) := C_2\vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3}$ with $C_2 > 0$ as in Lemma 3.2. In addition, let $\kappa_1(\vartheta) := c_{Z_0}(\alpha - \zeta_2 - \pi(\vartheta))$ and $\kappa_2(\vartheta) := c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta))$. To prove the first inequality, note that

$$\begin{aligned}
&\mathrm{P}(\{T \leqslant c_W(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \\
&\leqslant_{(1)} \mathrm{P}(\kappa_1(\vartheta) - 2\zeta_1 < T_0 \leqslant \kappa_2(\vartheta) + 2\zeta_1) + 2\mathrm{P}(\Delta > \vartheta) + 3\zeta_2 \\
&\leqslant_{(2)} \mathrm{P}(\kappa_1(\vartheta) - 2\zeta_1 < Z_0 \leqslant \kappa_2(\vartheta) + 2\zeta_1) + 2\mathrm{P}(\Delta > \vartheta) + 2\rho + 3\zeta_2 \\
&\leqslant_{(3)} 2\pi(\vartheta) + 2\mathrm{P}(\Delta > \vartheta) + 2\rho + C_3\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} + 5\zeta_2
\end{aligned}$$

where $C_3 > 0$ depends on c_1 and C_1 only and where (1) follows from equation (14) and Lemmas 3.2 and 3.3, (2) follows from the definition of ρ, and (3) follows from Lemma 2.1 and the fact that Z_0 has no point masses. The first inequality follows. The second inequality follows from the first inequality and the definition of ρ. ∎

C.5. Proof of Corollary 3.1. Since conditions (E.3) and (E.4) are special cases of (E.1) and (E.2), it suffices to prove the result under conditions (E.1) and (E.2) only. The proof of this corollary relies on:

Lemma C.1. *Recall conditions (E.1)-(E.2) in Section 2.1. Then*

$$\mathrm{E}[\Delta] \leqslant C \times \begin{cases} \sqrt{\frac{B_n^2 \log p}{n}} \bigvee \frac{B_n^2(\log(pn))^2(\log p)}{n}, & \textit{under (E.1)}, \\ \sqrt{\frac{B_n^2 \log p}{n}} \bigvee \frac{B_n^2(\log p)}{\sqrt{n}}, & \textit{under (E.2)}, \end{cases}$$

where $C > 0$ depends only on c_1 and C_1 that appear in (E.1)-(E.2).

Proof. By Lemma A.1 and Hölder's inequality, we have

$$\mathrm{E}[\Delta] \lesssim M_4^2\sqrt{(\log p)/n} + (\mathrm{E}[\max_{i,j}|x_{ij}|^4])^{1/2}(\log p)/n.$$

The conclusion of the lemma follows from elementary calculations with help of Lemma 2.2.2 in [30]. ∎

Proof of Corollary 3.1. To prove the first inequality, we make use of Theorem 3.2. Let $c > 0$ and $C > 0$ denote generic constants depending only on c_1, C_1, c_2, C_2, and their values may change from place to place. By Corollary 2.1, in all cases, $\rho \leqslant Cn^{-c}$. Moreover, $\zeta_1\sqrt{\log p} \leqslant C_2 n^{-c_2}$ implies that $\zeta_1 \leqslant C_2 n^{-c_2}$ (recall $p \geqslant 3$), and hence $\zeta_1\sqrt{\log(p/\zeta_1)} \leqslant Cn^{-c}$. Also, $\zeta_2 \leqslant Cn^{-c}$ by assumption.

Let $\vartheta = \vartheta_n := (\mathrm{E}[\Delta])^{1/2}/\log p$. By Lemma C.1, $\mathrm{E}[\Delta](\log p)^2 \leqslant Cn^{-c}$. Therefore, $\pi(\vartheta) \leqslant Cn^{-c}$ (with possibly different $c, C > 0$). In addition,

by Markov's inequality, $\mathrm{P}(\Delta > \vartheta) \leqslant \mathrm{E}[\Delta]/\vartheta \leqslant Cn^{-c}$. Hence, by Theorem 3.2, the first inequality follows. The second inequality follows from the first inequality and the fact that $\rho \leqslant Cn^{-c}$ as shown above. ∎

ACKNOWLEDGMENTS

The authors would like to express their appreciation to L.H.Y. Chen, David Gamarnik, Qi-Man Shao, Vladimir Koltchinskii, Enno Mammen, Axel Munk, Steve Portnoy, Adrian Röllin, Azeem Shaikh, and Larry Wasserman for enlightening discussions. We thank the editors and referees for the comments of the highest quality that have lead to substantial improvements.

SUPPLEMENTARY MATERIAL

Supplement to "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors" (). This supplemental file contains the additional technical proofs, theoretical and simulation results.

REFERENCES

[1] Alquier, P. and Hebiri, M. (2011). Generalization of ℓ_1 constraints for high dimensional regression problems. *Statist. Probab. Lett.* **81** 1760-1765.

[2] Anderson, M. (2008). Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *JASA* **103** 1481-1495.

[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] Belloni, A. and Chernozhukov, V. (2009). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521-547.

[6] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.

[7] Bretagnolle, J. and Massart, P. (1989). Hungarian construction from the non asymptotic viewpoint. *Ann. Probab.* **17** 239-256.

[8] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods,Theory and Applications.* Springer.

[9] Candès, E.J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313-2351.

[6] Chatterjee, S. (2005a). A simple invariance theorem. arXiv:math/0508213.

[11] Chatterjee, S. (2005b). An error bound in the Sudakov-Fernique inequality. arXiv:math/0510424.

[12] Chen, L., Goldstein, L. and Shao, Q.-M. (2011). *Normal Approximation by Stein's Method.* Springer.

[13] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012a). Central limit theorems and multiplier bootstrap when p is much larger than n. arXiv:1212.6906.

[10] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012b). Gaussian approximation of suprema of empirical processes. arXiv:1212.6906.

[11] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012c). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807.

[16] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Supplement to "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors."

[17] Chetverikov, D. (2011). Adaptive test of conditional moment inequalities. arXiv:1201.0167.

[18] Chetverikov, D. (2012). Testing regression monotonicity in econometric models. arXiv:1212.6757.

[12] de la Peña, V., Lai, T. and Shao, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer.

[13] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

[21] Dumbgen, L. and Spokoiny, V. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124-152.

[22] 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.

[23] Frick, K., Marnitz, P. and Munk, A. (2012). Shape-constrained regularization by statistical multiresolution for inverse problems: asymptotic analysis. *Inverse Problems* **28** 065006.

[24] Gautier, E. and Tsybakov, A. (2011). High-dimensional istrumental variables regression and confidence sets. arXiv: 1105.2454.

[25] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.

[26] He, X., and Shao, Q-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis* **73** 120-135.

[27] 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.

[19] 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.

[29] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799-828.

[30] 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.

[31] Leadbetter, M., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer.

[32] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21** 255-285.

[33] Panchenko, D. (2013). *The Sherrington-Kirkpatrick Model*. Springer.

[34] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

[35] Praestgaard, J. and Wellner, J.A. (1993). Exchangeably weighted bootstraps of the general empirical processes. *Ann. Probab.* **21** 2053-2086.

[26] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.

[27] Röllin, A. (2011). Stein's method in high dimensions with applications, *Ann. Inst. H. Poincaré Probab. Statist.* **49**, 529-549.

[38] Romano, J., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Stat. Assoc.* **100** 94-108.

[39] Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech.*

J. **41** 463-501.

[28] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.

[29] Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians*. Springer.

[30] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

[43] Ye, F. and Zhang, C. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519-3540.

Supplement to "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors"

by V. Chernozhukov, D. Chetverikov, and K. Kato

MIT, UCLA, and University of Tokyo

Supplementary Material I

Deferred Proofs for Results from Main Text

APPENDIX D: DEFERRED PROOFS FOR SECTION 2

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

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

Claim (b). Observe that

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

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). This follows from the Cauchy-Schwarz inequality and the definition of $\varphi(u)$.

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

Lemma D.1 (Tail Bounds for 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 D.1. See [12], 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 D.1 and the union bound, 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 claim (c), for $u \geqslant u(\gamma)$, with probability at least $1-\gamma$, for all $1 \leqslant j \leqslant p$,

$$\begin{aligned}\Lambda_j = 4\sqrt{\bar{\mathrm{E}}[(x_{ij}-\tilde{x}_{ij})^2]} + \sqrt{\mathbb{E}_n[(\mathrm{E}[x_{ij}1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\}])^2]} \\ \leqslant 5\sqrt{\bar{\mathrm{E}}[x_{ij}^2]}\varphi(u).\end{aligned}$$

The last two assertions imply claim (d). ∎

D.2. Proof of Theorem 2.2. Since M_2 is bounded from below and above by positive constants, we may normalize $M_2 = 1$, without loss of generality. In this proof, let $C > 0$ denote a generic constant depending only on c_1 and C_1, and its value may change from place to place.

For given $\gamma \in (0,1)$, denote $\ell_n := \log(pn/\gamma) \geqslant 1$ and let

$$u_1 := n^{3/8}\ell_n^{-5/8}M_3^{3/4} \text{ and } u_2 := n^{3/8}\ell_n^{-5/8}M_4^{1/2}.$$

Define $u := u(\gamma) \vee u_1 \vee u_2$ and $\beta := \sqrt{n}/(2\sqrt{2}u)$. Then $u \geqslant u(\gamma)$ and the choice of β trivially obeys $2\sqrt{2}u\beta \leqslant \sqrt{n}$. So, by Theorem 2.1 and using the argument as that in the proof of Corollary I.1, for every $\psi > 0$ and any $\bar{\varphi}(u) \geqslant \varphi(u)$, we have

$$\begin{aligned}\rho \leqslant C\big[&n^{-1/2}(\psi^3+\psi^2\beta+\psi\beta^2)M_3^3 + (\psi^2+\psi\beta)\bar{\varphi}(u) \\ &+\psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} + (\beta^{-1}\log p + \psi^{-1})\sqrt{1\vee\log(p\psi)} + \gamma\big].\end{aligned} \tag{25}$$

Step 1. We claim that we can take $\bar{\varphi}(u) := CM_4^2/u$ for all $u > 0$. Since $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$, we have $1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\} \leqslant 1\{|x_{ij}| > c_1^{1/2}u\}$. Hence

$$\begin{aligned}\bar{\mathrm{E}}[x_{ij}^2 1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\}] &\leqslant \bar{\mathrm{E}}[x_{ij}^2 1\{|x_{ij}| > c_1^{1/2}u\}] \\ &\leqslant \bar{\mathrm{E}}[x_{ij}^4 1\{|x_{ij}| > c_1^{1/2}u\}]/(c_1u^2) \leqslant \bar{\mathrm{E}}[x_{ij}^4]/(c_1u^2) \leqslant M_4^4/(c_1u^2).\end{aligned}$$

This implies $\varphi_x(u) \leqslant CM_4^2/u$. For $\varphi_y(u)$, note that

$$\bar{\mathrm{E}}[y_{ij}^4] = \mathbb{E}_n[\mathrm{E}[y_{ij}^4]] = 3\mathbb{E}_n[(\mathrm{E}[y_{ij}^2])^2] = 3\mathbb{E}_n[(\mathrm{E}[x_{ij}^2])^2] \leqslant 3\mathbb{E}_n[\mathrm{E}[x_{ij}^4]] = 3\bar{\mathrm{E}}[x_{ij}^4],$$

and hence $\varphi_y(u) \leqslant CM_4^2/u$ as well. This implies the claim of this step.

Step 2. We shall bound the right side of (25) by suitably choosing ψ depending on the range of u. In order to set up this choice we define $u^\star$ by the following equation:

$$\bar{\varphi}(u^\star)n^{3/8}/(M_3^3\ell_n^{5/6})^{3/4} = 1.$$

We then take

$$\psi = \psi(u) := \begin{cases} n^{1/8}\ell_n^{-3/8}M_3^{-3/4} & \text{if } u \geqslant u^\star, \\ \ell_n^{-1/6}(\bar{\varphi}(u))^{-1/3} & \text{if } u < u^\star. \end{cases} \tag{26}$$

We note that for $u < u^\star$,

$$\psi(u) \leqslant \psi(u^\star) = n^{1/8}\ell_n^{-3/8}M_3^{-3/4}.$$

That is, when $u < u^\star$ the smoothing parameter ψ is smaller than when $u \geqslant u^\star$.

Using these choices of parameters β and ψ and elementary calculations (which will be done in Step 3 below), we conclude from (25) that whether $u < u^\star$ or $u \geqslant u^\star$,

$$\rho \leqslant C(n^{-1/2}u\ell_n^{3/2} + \gamma).$$

The bound in the theorem follows from this inequality.

Step 3. (Computation of the bound on ρ). Note that since $\rho \leqslant 1$, we only had to consider the case where $n^{-1/2}u\ell_n^{3/2} \leqslant 1$ since otherwise the inequality is trivial by taking, say, $C = 1$. Since $u_1 = n^{3/8}M_3^{3/4}/\ell_n^{5/8}$ and $u_2 = n^{3/8}M_4^{1/2}/\ell_n^{5/8}$, we have

$$\begin{aligned} &(\bar{\varphi}(u^\star))^{4/3} = n^{-1/2}\ell_n^{5/6}M_3^3, \\ &\bar{\varphi}(u_1) \leqslant Cn^{-3/8}\ell_n^{5/8}M_4^2/M_3^{3/4}, \\ &\bar{\varphi}(u_2) \leqslant Cn^{-3/8}\ell_n^{5/8}M_4^{3/2}. \end{aligned}$$

Also note that $\psi \leqslant n^{1/8}$, and so $1 \vee \log(p\psi) \lesssim \log(pn) \leqslant \ell_n$. Therefore,

$$\beta^{-1}\log p\sqrt{1 \vee \log(p\psi)} \lesssim \beta^{-1}\ell_n^{3/2} \lesssim n^{-1/2}u\ell_n^{3/2}.$$

In addition, note that $\beta \lesssim \sqrt{n}/u \leqslant \sqrt{n}/u_1 = n^{1/8}\ell_n^{5/8}M_3^{-3/4} =: \bar{\beta}$ and $\psi \leqslant \bar{\beta}$ under either case. This implies that $(\psi^3 + \psi^2\beta + \psi\beta^2) \lesssim \psi\bar{\beta}^2$ and $(\psi^2 + \psi\beta) \leqslant \psi\bar{\beta}$.

Using these inequalities, we can compute the bounds claimed above.

(a). Bounding ρ when $u \geqslant u^\star$. Then

$$n^{-1/2}(\psi^3+\psi^2\beta+\psi\beta^2)M_3^3 \lesssim n^{-1/2}\psi\bar\beta^2 M_3^3 \leqslant n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2};$$
$$(\psi^2+\psi\beta)\bar\varphi(u) \lesssim \psi\bar\beta\bar\varphi(u) \leqslant \psi\bar\beta\bar\varphi(u^\star) \leqslant n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2};$$
$$\psi\bar\varphi(u)\sqrt{\log(p/\gamma)} \leqslant \psi\bar\beta\bar\varphi(u)\sqrt{\ell_n}/\bar\beta \leqslant \psi\bar\beta\bar\varphi(u^\star) \leqslant n^{-1/2}u\ell_n^{3/2}; \text{ and}$$
$$\psi^{-1}\sqrt{\ell_n} \leqslant n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2};$$

where we have used Step 1 and the fact that

$$\sqrt{\ell_n}/\bar\beta = \ell_n^{-1/2}\psi^{-1} \leqslant n^{-1/8}\ell_n^{-1/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2} \leqslant 1.$$

The claimed bound on ρ now follows.

(b). Bounding ρ when $u < u^\star$. Since ψ is smaller than in case (a), by the calculations in Step (a)

$$n^{-1/2}(\psi^3+\psi^2\beta+\psi\beta^2)M_3^3/\sqrt{n} \lesssim n^{-1/2}u\ell_n^{3/2}.$$

Moreover, using definition of ψ, $u > u_2$, definition of u_2, we have

$$\psi\beta\bar\varphi(u) \leqslant \beta\bar\varphi(u)^{2/3}\ell_n^{-1/6} \leqslant \beta\bar\varphi(u_2)^{2/3}\ell_n^{-1/6} \leqslant n^{-1}\beta u_2^2\ell_n^{5/3-1/6} \lesssim n^{-1/2}u\ell_n^{3/2};$$
$$\psi^2\bar\varphi(u) \leqslant \bar\varphi(u)^{1/3}\ell_n^{-1/3} \leqslant \bar\varphi(u_2)^{1/3}\ell_n^{-1/3} \leqslant n^{-1/2}u_2\sqrt{\ell_n} \leqslant n^{-1/2}u\ell_n^{3/2}.$$

Analogously and using $n^{-1/2}u\ell_n^{3/2} \leqslant 1$, we have

$$\psi\bar\varphi(u)\sqrt{\log(p/\gamma)} \leqslant \bar\varphi(u)^{2/3}\ell_n^{1/3} \leqslant \bar\varphi(u_2)^{2/3}\ell_n^{1/3} \leqslant n^{-1}u_2^2\ell_n^2 \leqslant n^{-1/2}u\ell_n^{3/2}.$$
$$\psi^{-1}\sqrt{\ell_n} = \bar\varphi(u)^{1/3}\ell_n^{2/3} \leqslant n^{-1/2}u\ell_n^{3/2}.$$

This completes the proof. ∎

APPENDIX E: DEFERRED PROOFS FOR SECTION 4

E.1. Proof of Theorem 4.1. The proof proceeds in three steps. In the proof $(\widehat\beta,\lambda)$ denotes $(\widehat\beta^{(k)},\lambda^{(k)})$ with k either 0 or 1.

Step 1. Here we show that there exist some constants $c>0$ and $C>0$ (depending only c_1, C_1 and σ^2) such that

$$\mathrm{P}(T_0 \leqslant \lambda) \geqslant 1-\alpha-\nu_n, \tag{27}$$

with $\nu_n = Cn^{-c}$. 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 2.1-(ii) 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. The claim follows since $\lambda \geqslant c_{Z_0}(1-\alpha)$, which holds because $\lambda^{(1)} = c_{Z_0}(1-\alpha)$, and $\lambda^{(1)} \leqslant \lambda^{(0)} = c_0(1-\alpha) := \sigma\Phi^{-1}(1-\alpha/(2p))$ (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 $\geqslant 1-\alpha-\nu_n$, $\widehat{\delta} = \widehat{\beta} - \beta$ 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 implies $\sqrt{n}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]| \leqslant T_0 + \lambda$. The claim follows from Step 1.

Step 3. By Step 1, with probability $\geqslant 1-\alpha-\nu_n$, the true value β obeys the constraint in optimization problem (16) in the main text, in which case by definition of $\widehat{\beta}$, $\|\widehat{\beta}\|_{\ell_1} \leqslant \|\beta\|_{\ell_1}$. 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 with the same probability,

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

Combining this inequality with Step 2 gives the claim of the theorem. ∎

E.2. Proof of Theorem 4.2. 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 $\lambda = \lambda^{(0)} := c_0(1-1/n)$, where 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 $\varepsilon_i - \widehat{\varepsilon}_i = z_i'(\widehat{\beta}^{(0)} - \beta)$ gives

$$\max_{1\leqslant j\leqslant p}\left(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2]\right)^{1/2} \leqslant B_n(\mathbb{E}_n[z_i'(\widehat{\beta}^{(0)} - \beta)]^2)^{1/2} \leqslant B_n\|\widehat{\beta}^{(0)} - \beta\|_{\mathrm{pr}}.$$

The claim follows from Step 0.

Step 2. In this step, we apply Corollary 3.1-(ii) to

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

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

$$\mathrm{P}(T_0 \leqslant c_W(1-\alpha)) \geqslant 1 - \alpha - \varrho_n. \tag{28}$$

To show applicability of Corollary 3.1-(ii), we note that for any $\zeta_1 > 0$,

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

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 take 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 the conditions of Corollary 3.1-(ii) with so defined ζ_1 and $\zeta_2 = \varrho_n \vee ((\iota_n^2 \log p)^{1/2}/\zeta_1)$ are satisfied, and hence application of the corollary gives that uniformly in $\alpha \in (0,1)$,

$$|\mathrm{P}(T_0 \leqslant c_W(1-\alpha)) - 1 - \alpha| \leqslant \varrho_n, \tag{29}$$

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 3.3 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 3.2 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 these bounds gives the claim of this step.

Step 4. Given (28), the rest of the proof is identical to Steps 2-3 in the proof of Theorem 4.1 with $\lambda = c_W(1-\alpha)$. The result follows for $\nu_n = 2\varrho_n$.■

APPENDIX F: DEFERRED PROOFS FOR SECTION 5

F.1. Proof of Theorem 5.1. The multiplier bootstrap critical values $c_{1-\alpha,w}$ clearly satisfy $c_{1-\alpha,w} \leqslant c_{1-\alpha,w'}$ whenever $w \subset w'$, so inequality (22) in the main text is satisfied. Therefore, it suffices to prove (23) in the main text.

Let w denote the set of true null hypotheses. Then for any $j \in w$,

$$t_j = \sqrt{n}(\widehat{\beta}_j - \beta_{0j}) \leqslant \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} + r_{nj}.$$

Therefore, for all $j \in w$, we can and will assume that $\beta_j = \beta_{0j}$.

For $w \subset \mathcal{W} = \{1, \ldots, p\}$, define

$$T := T(w) := \max_{j \in w} \sqrt{n}(\widehat{\beta}_j - \beta_{0j}),\ W := W(w) := \max_{j \in w} \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{x}_{ij} e_i.$$

In addition, define

$$T_0 := T_0(w) := \max_{j \in w} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij},\ W_0 := W_0(w) := \max_{j \in w} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_i.$$

To prove (23), we will apply Corollary 3.1. By assumption, either (i) or (ii) of Corollary 3.1 holds. Therefore, it remains to verify conditions in equations (14) and (15) in the main text with $\zeta_1 \sqrt{\log p} + \zeta_2 \leqslant C n^{-c}$ for some $c > 0$ and $C > 0$ uniformly over all $w \subset \mathcal{W}$.

Set $\zeta_1 = (C/2) n^{-c} / \sqrt{\log p}$ and $\zeta_2 = (C/2) n^{-c}$ for sufficiently small $c > 0$ and large $C > 0$ depending on c_1, C_1, c_2, and C_2 only. Note that $\zeta_1 \sqrt{\log p} + \zeta_2 \leqslant C n^{-c}$. Also note that

$$|T - T_0| \leqslant \max_{1 \leqslant j \leqslant p} |r_{nj}| = \Delta_1$$

for all $w \subset \mathcal{W}$. Therefore, it follows from assumption (i) that $\mathrm{P}(|T - T_0| > \zeta_1) < \zeta_2$ for all $w \subset \mathcal{W}$, i.e. condition in equation (14) holds uniformly over all $w \subset \mathcal{W}$. Further, note that $\sum_{i=1}^n (\widehat{x}_{ij} - x_{ij}) e_i / \sqrt{n}$ conditional on $(x_i)_{i=1}^n$ and $(\widehat{x}_i)_{i=1}^n$ is distributed as $N(0, \mathbb{E}_n[(\widehat{x}_{ij} - x_{ij})^2])$ random variable and

$$|W - W_0| \leqslant \max_{1 \leqslant j \leqslant p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\widehat{x}_{ij} - x_{ij}) e_i \right|$$

for all $w \subset \mathcal{W}$. Therefore, $\mathrm{E}_e[|W - W_0|] \leqslant (C/2)\sqrt{\Delta_2 \log p}$, and so it follows from Borell inequality and assumption (ii) that $\mathrm{P}(\mathrm{P}_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2$ for all $w \subset \mathcal{W}$, i.e. condition in equation (15) holds uniformly over all $w \subset \mathcal{W}$. This completes the proof by applying Corollary 3.1. ∎

APPENDIX G: MONTE CARLO EXPERIMENTS IN SUPPORT OF SECTION 4

In this section, we present results of Monte Carlo simulations that illustrate our theoretical results on Dantzig selector given in Section 4. We

consider Gaussian and Non-Gaussian noise with homoscedasticity and heteroscedasticity. We study 3 types of Dantzig selector depending on the choice of the penalty level: canonical, ideal (based on gaussian approximation, GAR), and multiplier bootstrap (MB).

We consider the following regression model:

$$y_i = z_i'\beta + \varepsilon_i,$$

where observations are independent across i, y_i is a scalar dependent variable, z_i is a p-dimensional vector of covariates, and ε_i is noise. The first component of z_i equals 1 in all experiments (an intercept). Other $p-1$ components are simulated as follows: first, we simulate a vector $w_i \in \mathbb{R}^{p-1}$ from the Gaussian distribution with zero mean so that $\mathrm{E}[w_{ij}^2] = 1$ for all $1 \leqslant j \leqslant p-1$ and $\mathrm{E}[w_{ij}w_{ik}] = \rho$ for all $1 \leqslant j, k \leqslant p-1$ with $j \neq k$; second, we set $z_{ij+1} = w_{ij}/(\mathbb{E}_n[w_{ij}^2])^{1/2}$ (equicorrelated design). Depending on the experiment, we set $\rho = 0$, 0.5, 0.9, or 0.99. We simulate $\epsilon_i = \sigma_0\sigma(z_i)e_i$ where depending on the experiment, $\sigma_0 = 0.5$ or 1.0 and e_i is taken either from $N(0,1)$ distribution (Gaussian noise) or from t-distribution with 5 degrees of freedom normalized to have variance 1 (Non-Gaussian noise). To investigate the effect of heteroscedasticity on the properties of different estimators, we set

$$\sigma(z_i) = \frac{2\exp(\gamma z_{i2})}{1+\exp(\gamma z_{i2})}$$

where γ is either 0 (homoscedastic case) or 1 (heteroscedastic case).

Tables 1 and 2 present results on prediction error of Dantzig selector for the case of Non-Gaussian and Gaussian noise, respectively. Prediction error is defined as

$$\|\widehat{\beta} - \beta\|_{pr} = \sqrt{\mathbb{E}_n[z_i'(\widehat{\beta} - \beta)]}$$

where $\widehat{\beta}$ is the Dantzig selector; see Section 4 for the definition of the Dantzig selector. Recall that implementing the Dantzig selector requires selecting the penalty level λ. Both tables show results for 3 different choices of the penalty level. Canonical penalty is $\lambda = \bar{\sigma}\Phi^{-1}(1-\alpha/(2p))$ where $\bar{\sigma} = \sigma_0(1 + I\{|\gamma| > 0\})$, the upper bound on the variance of ε_i's. Ideal (based on gaussian approximation, GAR) penalty is $\lambda = c_{Z_0}(1-\alpha)$, the conditional $(1-\alpha)$ quantile of Z_0 given $(z_i)_{i=1}^n$ where

$$Z_0 = \sqrt{n}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}\sigma_0\sigma(z_i)e_i]|$$

where $e_i \sim N(0,1)$ independently across i. Finally, multiplier bootstrap (MB) penalty is defined as follows. First, we calculate the Dantzig selector with the canonical choice of the penalty level, $\widehat{\beta}$, and select regressors corresponding to non-zero components of $\widehat{\beta}$. Second, we run the OLS regression

TABLE 1
Results of Monte Carlo experiments for prediction error. Non-Gaussian noise.

Distribution ε	σ_0	ρ	Method		
			Canonical	GAR	MB
Homoscedastic	0.5	0.00	0.224	0.207	0.208
		0.50	0.390	0.353	0.379
		0.90	0.352	0.317	0.340
		0.99	0.107	0.057	0.058
	1.0	0.00	0.648	0.604	0.674
		0.50	0.695	0.643	0.644
		0.90	0.538	0.406	0.412
		0.99	0.348	0.139	0.137
Heteroscedastic	0.5	0.00	0.656	0.252	0.393
		0.50	0.660	0.407	0.469
		0.90	0.516	0.326	0.342
		0.99	0.339	0.066	0.064
	1.0	0.00	1.588	0.661	0.909
		0.50	1.590	0.722	0.755
		0.90	1.336	0.445	0.454
		0.99	1.219	0.153	0.156

TABLE 2
Results of Monte Carlo experiments for prediction error. Gaussian noise.

Distribution ε	σ_0	ρ	Method		
			Canonical	GAR	MB
Homoscedastic	0.5	0.00	0.229	0.211	0.210
		0.50	0.421	0.386	0.417
		0.90	0.365	0.315	0.350
		0.99	0.109	0.059	0.059
	1.0	0.00	0.663	0.618	0.671
		0.50	0.679	0.627	0.624
		0.90	0.554	0.424	0.429
		0.99	0.326	0.127	0.120
Heteroscedastic	0.5	0.00	0.674	0.257	0.395
		0.50	0.643	0.412	0.451
		0.90	0.503	0.310	0.324
		0.99	0.319	0.060	0.059
	1.0	0.00	1.690	0.708	0.976
		0.50	1.537	0.665	0.679
		0.90	1.334	0.439	0.452
		0.99	1.189	0.155	0.148

of y_i on the set of selected regressors, and take residuals from this regression, $(\widehat{\epsilon}_i)_{i=1}^n$. Then the multiplier bootstrap penalty level is $\lambda = c_W(1-\alpha)$, the conditional $(1-\alpha)$ quantile of W given $(z_i, \widehat{\varepsilon}_i)_{i=1}^n$ where

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

where $e_i \sim N(0,1)$ independently across i.

The results show that the GAR penalty always yields smaller prediction error than that of the canonical penalty. Moreover, as predicted by the theory, GAR penalty works especially good in comparison with the canonical penalty in heteroscedastic case and/or in the case with high correlation between regressors (high ρ). In addition, in most cases, the results for the MB penalty are similar to those for the GAR penalty. In particular, the MB penalty in most cases yields smaller prediction error than that of the canonical penalty. Finally, the GAR penalty in most cases is slightly better than the MB penalty. Note, however, that when heteroscedasticity function $\sigma(z_i)$ is unknown, the GAR penalty becomes infeasible but the MB penalty is feasible given that the upper bound on the variance of ε_i's exists.

Supplementary Material II

Additional Results and Discussions

APPENDIX H: A NOTE ON SLEPIAN-STEIN TYPE METHODS FOR NORMAL APPROXIMATIONS

To keep the notation simple, consider a random vector X in $\mathbb{R}^p$ and a standard normal vector Y in $\mathbb{R}^p$. We are interested in bounding

$$\mathrm{E}[g(X)] - \mathrm{E}[g(Y)],$$

over some collection of test functions $g \in \mathcal{G}$. Without loss of generality, suppose that Y and X are independent.

Consider Stein's partial differential equation:

$$g(x) - \mathrm{E}[g(Y)] = \triangle h(x) - x'\nabla h(x)$$

where $\triangle h(X)$ and $\nabla h(X)$ refer to the Laplacian and the gradient of $h(X)$. It is well known, e.g. [14] and [8], that an explicit solution for h in this equation is given by

$$h(x) := -\int_0^1 \frac{1}{2t}\left[\mathrm{E}[g(\sqrt{t}x + \sqrt{1-t}Y)] - \mathrm{E}[g(Y)]\right] dt,$$

so that

$$\mathrm{E}[g(X)] - \mathrm{E}[g(Y)] = \mathrm{E}[\triangle h(X) - X'\nabla h(X)].$$

The Stein type method for normal approximation bounds the right side for $g \in \mathcal{G}$.

Next, let us consider the Slepian smart path interpolation:

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

Then we have

$$\mathrm{E}[g(X)] - \mathrm{E}[g(Y)] = \mathrm{E}\left[\int_0^1 \frac{1}{2}\nabla g(Z(t))'\left(\frac{X}{\sqrt{t}} - \frac{Y}{\sqrt{1-t}}\right)\right] dt.$$

The Slepian type method, as used in our paper, bounds the right side for $g \in \mathcal{G}$. We also refer the reader to [27] for a related discussion and interesting results (see in particular Lemma 2.1 in [27]).

Elementary calculations and integration by parts yield the following observation.

LEMMA H.1. *Suppose that $g : \mathbb{R}^p \to \mathbb{R}$ is a C^2-function with uniformly bounded derivatives up to order two. Then*

$$I := \mathrm{E}\left[\int_0^1 \frac{1}{2}\nabla g(Z(t))'\left(\frac{X}{\sqrt{t}}\right)\right] = -\mathrm{E}[X'\nabla h(X)]$$

and

$$II := \mathrm{E}\left[\int_0^1 \frac{1}{2}\nabla g(Z(t))'\left(\frac{Y}{\sqrt{1-t}}\right)\right] = -\mathrm{E}[\triangle h(X)].$$

Hence the Slepian and Stein methods both show that difference between I and II is small or approaches zero under suitable conditions on X; therefore, they are very similar in spirit, if not identical. The details of treating terms may be different from application to application; see more on this in [27].

PROOF OF LEMMA H.1. By definition of h, we have

$$-\mathrm{E}[X'\nabla h(X)] = \mathrm{E}\left[X'\int_0^1 \frac{1}{2t}\nabla g(Z(t))\sqrt{t}dt\right] = \mathrm{E}\left[\int_0^1 \nabla g(Z(t))'\frac{X}{2\sqrt{t}}dt\right].$$

On the other hand, by definition of h and Stein's identity (Lemma H.2),

$$-\mathrm{E}[\triangle h(X)] = \mathrm{E}\left[\frac{1}{2}\int_0^1 \triangle g(Z(t))dt\right] = \mathrm{E}\left[\frac{1}{2}\int_0^1 \nabla g(Z(t))'\left(\frac{Y}{\sqrt{1-t}}\right)dt\right].$$

This completes the proof. ∎

LEMMA H.2 (Stein's identity). *Let $W = (W_1, \ldots, W_p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a C^1-function such that $\mathrm{E}[|\partial_j f(W)|] < \infty$ for all $1 \leqslant j \leqslant p$. Then for every $1 \leqslant j \leqslant p$,*

$$\mathrm{E}[W_j f(W)] = \sum_{k=1}^p \mathrm{E}[W_j W_k]\mathrm{E}[\partial_k f(W)].$$

PROOF OF LEMMA H.2. See Section A.6 of [29], and also [28]. ∎

APPENDIX I: A SIMPLE GAUSSIAN APPROXIMATION RESULT

This section can be helpful to the reader wishing to see how Slepian-Stein methods can be used to prove a simple Gaussian approximation (whose applicability is limited however.) We start with the following elementary lemma.

LEMMA I.1 (A Simple Comparison of Gaussian to Non-Gaussian Maxima). *For every $g \in C_b^3(\mathbb{R})$ and $\beta > 0$,*

$$|\mathrm{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[S_i^3],$$

and hence

$$|\mathrm{E}[g(T_0) - g(Z_0)]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[S_i^3] + \beta^{-1}G_1 \log p.$$

The optimal value of the last bound is given by taking the minimum over β. We postpone choices of β to the proof of the subsequent corollary, leaving ourselves more flexibility in optimizing bounds in the corollary.

COMMENT I.1. The bound above per se seems new, though it is merely a simple extension of results in [6], who obtained the bound for the case with X having a special structure like in our example (E.4), related to spin glasses, using classical Lindeberg's method. We give a proof using a variant of Slepian-Stein method, since this is the tool we end up using to prove our main results, as the Lindeberg's method, in its pure form, did not yield the same sharp results. Our proof is related but rather different in details from the more abstract/general arguments based on Stein triplets given in [27] (Lemma 2.1), but given for the special case of data $(x_i)_{i=1}^n$ with coordinates x_i's that $\mathbb{R}$-valued, in contrast to the $\mathbb{R}^p$-valued case treated here. [27] reanalyzed [6]'s setup under local dependence and gave a number of other interesting applications. ■

The next result states a bound on the Kolmogorov distance between distributions of T_0 and Z_0. The result follows from Lemma I.1 and the anti-concentration inequality for maxima of Gaussian random variables stated in Lemma 2.1. Note that this result was not included in either [6] or [27] for the cases that they have analyzed.

COROLLARY I.1 (**A Simple Gaussian Approximation**). *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 there exists a constant $C > 0$ depending only on c_1 and C_1 such that*

$$\rho := \sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)| \leqslant C(n^{-1}(\log(pn))^7)^{1/8}(\bar{\mathrm{E}}[S_i^3])^{1/4}.$$

Theorem I.1 and Corollary I.1 imply that the error of approximating the maximum coordinate in the sum of independent random vectors by its Gaussian analogue depends on p (possibly) only through $\log p$. This is the

main qualitative feature of all the results in this paper. Both Lemma I.1 and Corollary I.1 and all the results in this paper do not limit the dependence among the coordinates in x_i.

While Lemma I.1 and Corollary I.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 $\bar{\mathrm{E}}[S_i^3]$. When $\bar{\mathrm{E}}[S_i^3]$ increases with n, for example when $|x_{ij}| \leqslant B_n$ for all i and j and B_n grows with n, the simple bound above may be too poor, and can be improved considerably using several inputs. We derive in Theorem 2.1 in the main text a bound that can be much better in the latter scenario. The improvement there comes at a cost of more involved statements and proofs.

Proof of Lemma I.1. Without loss of generality, we are assuming that sequences $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent. For $t \in [0,1]$, we consider the Slepian interpolation between Y and X:

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

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

$$Z^{(i)}(t) := Z(t) - Z_i(t).$$

Let $\Psi(t) = \mathrm{E}[m(Z(t))]$ for $m := g \circ F_\beta$. Then by Taylor's theorem,

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

where

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

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

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

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

Note that random vector $Z^{(i)}(t)$ is independent of $(\dot{Z}_{ij}(t), Z_{ij}(t))$, and $\mathrm{E}[\dot{Z}_{ij}(t)] = 0$. Hence we have $I = 0$; moreover, since $\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)] = n^{-1}\mathrm{E}[x_{ij}x_{ik} - y_{ij}y_{ik}] = 0$ by construction of $(y_i)_{i=1}^n$, we also have $II = 0$. Consider the third term III. We have that

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

where (1) follows from $|\partial_j\partial_k\partial_l m(Z^{(i)}(t) + \tau Z_i(t))| \leqslant U_{jkl}(Z^{(i)}(t) + \tau Z_i(t)) \lesssim (G_3 + G_2\beta + G_1\beta^2)$ holding by Lemma A.5, and (2) is shown below. The first claim of the theorem now follows. The second claim follows directly from property (8) in the main text of the smooth max function.

It remains to show (2). Define $\omega(t) = 1/(\sqrt{t} \wedge \sqrt{1-t})$ and note,

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

where the first inequality follows from Hölder's inequality, and the second from the fact that $|\dot{Z}_{ij}(t)/\omega(t)| \leqslant (|x_{ij}| + |y_{ij}|)/\sqrt{n}$, $|Z_{ij}(t)| \leqslant (|x_{ij}| + |y_{ij}|)/\sqrt{n}$. Finally we note that $\int_0^1 \omega(t)dt \lesssim 1$, so inequality (2) follows. This completes the overall proof. ∎

Proof of Corollary I.1. In this proof, let $C > 0$ denote a generic constant depending only on c_1 and C_1, and its value may change from place to place. For $\beta > 0$, define $e_\beta := \beta^{-1}\log p$. Recall that $S_i := \max_{1\leqslant j\leqslant p}(|x_{ij}| + |y_{ij}|)$. Consider and fix a $C_b^3(\mathbb{R})$-function $g_0 : \mathbb{R} \to [0,1]$ 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, $G_0 = 1$, $G_1 \lesssim \psi$, $G_2 \lesssim \psi^2$ and $G_3 \lesssim \psi^3$.

Observe now that

$$\begin{aligned}\mathrm{P}(T_0 \leqslant t) &\leqslant \mathrm{P}(F_\beta(X) \leqslant t + e_\beta) \leqslant \mathrm{E}[g(F_\beta(X))]\\ &\leqslant \mathrm{E}[g(F_\beta(Y))] + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3])\\ &\leqslant \mathrm{P}(F_\beta(Y) \leqslant t + e_\beta + \psi^{-1}) + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3])\\ &\leqslant \mathrm{P}(Z_0 \leqslant t + e_\beta + \psi^{-1}) + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3]).\end{aligned}$$

where the first inequality follows from (8), the second from construction of g, the third from Theorem I.1, and the fourth from construction of g, and the last from (8). The remaining step is to compare $\mathrm{P}(Z_0 \leqslant t + e_\beta + \psi^{-1})$ with $\mathrm{P}(Z_0 \leqslant t)$ and this is where Lemma 2.1 plays its role. By Lemma 2.1,

$$\mathrm{P}(Z_0 \leqslant t + e_\beta + \psi^{-1}) - \mathrm{P}(Z_0 \leqslant t) \leqslant C(e_\beta + \psi^{-1})\sqrt{1 \vee \log(p\psi)}.$$

by which we have

$$\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t) \leqslant C[(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3]) + (e_\beta + \psi^{-1})\sqrt{1 \vee \log(p\psi)}].$$

We have to minimize the right side with respect to β and ψ. It is reasonable to choose β in such a way that e_β and ψ^{-1} are balanced, i.e., $\beta = \psi \log p$. With this β, the bracket on the right side is bounded from above by

$$C[\psi^3(\log p)^2(n^{-1/2}\bar{\mathrm{E}}[S_i^3]) + \psi^{-1}\sqrt{1 \vee \log(p\psi)}],$$

which is approximately minimized by $\psi = (\log p)^{-3/8}(n^{-1/2}\bar{\mathrm{E}}[S_i^3])^{-1/4}$. With this ψ, $\psi \leqslant (n^{-1/2}\bar{\mathrm{E}}[S_i^3])^{-1/4} \leqslant Cn^{1/8}$ (recall that $p \geqslant 3$), and hence $\log(p\psi) \leqslant C\log(pn)$. Therefore,

$$\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t) \leqslant C(n^{-1/2}\bar{\mathrm{E}}[S_i^3])^{1/4}(\log(pn))^{7/8}.$$

This gives one half of the claim. The other half follows similarly. ∎

APPENDIX J: GAUSSIAN APPROXIMATION AND MULTIPLIER BOOTSRAP RESULTS, ALLOWING FOR LOW VARIANCES

The purpose of this section is to provide results without an assumption that $\bar{\mathrm{E}}[x_{ij}^2] > c$ for all $1 \leqslant j \leqslant p$ and some constant $c > 0$.

J.1. Gaussian Approximation Results. In this subsection, we use the same setup and notation as those in Section 2. In particular, $x_1, \ldots, x_n$ is a sequence of independent centered random vectors in $\mathbb{R}^p$, $y_1, \ldots, y_n$ is a sequence of independent centered Gaussian random vectors such that $\mathrm{E}[y_i y_i'] = \mathrm{E}[x_i x_i']$, $T_0 = \max_{1 \leqslant j \leqslant p} X_j$ where $X = \sum_{i=1}^n x_i/\sqrt{n}$, $Z_0 = \max_{1 \leqslant j \leqslant p} Y_j$ where $Y = \sum_{i=1}^n y_i/\sqrt{n}$, and

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

In addition, denote

$$M_{k,2} := \max_{1 \leqslant j \leqslant p} \frac{\bar{\mathrm{E}}[|x_{ij}|^k]^{1/k}}{\bar{\mathrm{E}}[x_{ij}^2]^{1/2}} \text{ and } \ell_n := \log(pn/\gamma).$$

We will impose the following condition:

(SM) There exists $J \subset \{1, \dots, p\}$ such that $|J| \geqslant \nu p$ and for all $(j,k) \in J \times J$ with $j \neq k$, $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$ and $|\bar{\mathrm{E}}[x_{ij}x_{ik}]| \leqslant (1-\nu')(\bar{\mathrm{E}}[x_{ij}^2]\bar{\mathrm{E}}[x_{ik}^2])^{1/2}$ for some strictly positive constants ν, ν', and c_1 independent of n.

THEOREM J.1. *Suppose that condition (SM) holds. In addition, suppose that there is some constant $C_1 > 0$ such that $\bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for $1 \leqslant j \leqslant p$. Then for every $\gamma \in (0,1)$,*

$$\rho \leqslant C\left\{n^{-1/8}(M_3^{3/4} \vee M_{4,2}^{1/2})\ell_n^{7/8} + n^{-1/2}\ell_n^{3/2}u(\gamma) + p^{-c}\ell_n^{1/2} + \gamma\right\},$$

where $c, C > 0$ are constants that depend only on ν, ν', c_1 and C_1.

Theorem J.1 has the following applications. Let $C_1 > 0$ be some constant that is independent of n, and let $B_n \geqslant 1$ be a sequence of constants. We allow for the case where $B_n \to \infty$ as $n \to \infty$. We will assume that one of the following conditions is satisfied *uniformly in* $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$:

(E.5) $\bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ and $\max_{k=1,2} M_{k+2,2}^{k+2}/B_n^k + \mathrm{E}[\exp(|x_{ij}|/B_n)] \leqslant 2$;

(E.6) $\bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ and $\max_{k=1,2} M_{k+2,2}^{k+2}/B_n^k + \mathrm{E}[(\max_{1\leqslant j\leqslant p}|x_{ij}|/B_n)^4] \leqslant 2$.

COROLLARY J.1. *Suppose that there exist constants $c_2 > 0$ and $C_2 > 0$ such that one of the following conditions is satisfied: (i) (E.5) holds and $B_n^2(\log(pn))^7/n \leqslant C_2 n^{-c_2}$ or (ii) (E.6) holds and $B_n^4(\log(pn))^7/n \leqslant C_2 n^{-c_2}$. In addition, suppose that condition (SM) holds and $p \geqslant C_3 n^{c_3}$ for some constants $c_3 > 0$ and $C_3 > 0$. Then there exist constants $c > 0$ and $C > 0$ depending only on $\nu, \nu', c_1, C_1, c_2, C_2, c_3$, and C_3 such that*

$$\rho \leqslant Cn^{-c}.$$

J.2. Multiplier Bootstrap Results. In this subsection, we use the same setup and notation as those in Section 3. In particular, in addition to the notation used above, we assume that random variables T and W satisfy conditions (14) and (15) in the main text, respectively, where $\zeta_1 \geqslant 0$ and $\zeta_2 \geqslant 0$ depend on n and where W_0 appearing in condition (15) is defined in equation (13) in the main text. Recall that $\Delta = \max_{1\leqslant j\leqslant p}|\mathbb{E}_n[x_{ij}] - \bar{\mathrm{E}}[x_{ij}]|$.

THEOREM J.2. *Suppose that there is some constant $C_1 > 0$ such that $\bar{\sigma} := \max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. In addition, suppose that condition (SM) holds. Moreover, suppose that conditions (14) and (15) are satisfied. Then for every $\vartheta > 0$,*

$$\begin{aligned}\rho_\ominus &:= \sup_{\alpha\in(0,1)} \mathrm{P}(\{T \leqslant c_W(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \\ &\leqslant 2(\rho + \pi(\vartheta) + \mathrm{P}(\Delta > \vartheta)) + C(\zeta_1 \vee p^{-c})\sqrt{1 \vee \log(p/\zeta_1)} + 5\zeta_2,\end{aligned}$$

where

$$\pi(\vartheta) := C\vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3} + Cp^{-c}\sqrt{1 \vee \log(p/\vartheta)}$$

and $c, C > 0$ *depend only on* ν, ν', c_1 *and* C_1. *In addition,*

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant \rho_\ominus + \rho.$$

COROLLARY J.2. *Suppose that there exist constants* $c_2, C_2 > 0$ *such that conditions (14) and (15) hold with* $\zeta_1\sqrt{\log p} + \zeta_2 \leqslant C_2 n^{-c_2}$. *Moreover, suppose that one of the following conditions is satisfied: (i) (E.5) holds and* $B_n^2(\log(pn))^7/n \leqslant C_2 n^{-c_2}$ *or (ii) (E.6) holds and* $B_n^4(\log(pn))^7/n \leqslant C_2 n^{-c_2}$. *Finally, suppose that condition (SM) holds and* $p \geqslant C_3 n^{c_3}$ *for some constants* $c_3 > 0$ *and* $C_3 > 0$. *Then there exist constants* $c > 0$ *and* $C > 0$ *depending only on* $\nu, \nu', c_1, C_1, c_2, C_2, c_3,$ *and* C_3 *such that*

$$\rho_\ominus = \sup_{\alpha \in (0,1)} \mathrm{P}(\{T \leqslant c_W(\alpha)\} \ominus \{T_0 \leqslant c_{Z_0}(\alpha)\}) \leqslant Cn^{-c}.$$

In addition, $\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant \rho_\ominus + \rho \leqslant Cn^{-c}$.

The proofs rely on the following auxiliary lemmas, whose proofs will be given below.

LEMMA J.1. *(a) Let* $Y_1, \ldots, Y_p$ *be jointly Gaussian random variables with* $\mathrm{E}[Y_j] = 0$ *and* $\sigma_j^2 := \mathrm{E}[Y_j^2]$ *for all* $1 \leqslant j \leqslant p$. *Let* $b_p := \mathrm{E}[\max_{1 \leqslant j \leqslant p} Y_j]$ *and* $\bar{\sigma} = \max_{1 \leqslant j \leqslant p} \sigma_j > 0$. *Assume that* $b_p \geqslant c_1\sqrt{\log p}$ *for some* $c_1 > 0$. *Then for every* $\varsigma > 0$,

$$\sup_{z \in \mathbb{R}} \mathrm{P}\left(|\max_{1 \leqslant j \leqslant p} Y_j - z| \leqslant \varsigma\right) \leqslant C(\varsigma \vee p^{-c})\left(b_p + \sqrt{1 \vee \log(\bar{\sigma}/\varsigma)}\right) \tag{30}$$

where $c, C > 0$ *are some constants depending only on* c_1 *and* $\bar{\sigma}$. *(b) Furthermore, the worst case bound is obtained by bounding* b_p *by* $\bar{\sigma}\sqrt{2\log p}$.

LEMMA J.2. *Let* V *and* Y *be centered Gaussian random vectors in* $\mathbb{R}^p$ *with covariance matrices* Σ^V *and* Σ^Y, *respectively. Let* $\Delta_0 := \max_{1 \leqslant j,k \leqslant p} |\Sigma_{jk}^V - \Sigma_{jk}^Y|$. *Suppose that there are some constants* $0 < c_1 < C_1$ *such that* $\bar{\sigma} := \max_{1 \leqslant j \leqslant p} \mathrm{E}[Y_j^2] \leqslant C_1$ *for all* $1 \leqslant j \leqslant p$ *and* $b_p := \mathrm{E}[\max_{1 \leqslant j \leqslant p} Y_j] \geqslant c_1\sqrt{\log p}$. *Then there exist constants* $c > 0$ *and* $C > 0$ *depending only on* c_1 *and* C_1 *such that*

$$\sup_{t \in \mathbb{R}} \left|\mathrm{P}\left(\max_{1 \leqslant j \leqslant p} V_j \leqslant t\right) - \mathrm{P}\left(\max_{1 \leqslant j \leqslant p} Y_j \leqslant t\right)\right| \leqslant C\Delta_0^{1/3}(1 \vee \log(p/\Delta_0))^{2/3} + Cp^{-c}\sqrt{1 \vee \log(p/\Delta_0)}.$$

Proof of Theorem J.1. It follows from Theorem 2.3.16 in [13] that condition (SM) implies that $\mathrm{E}[Z_0] \geqslant c\sqrt{\log p}$ for some $c > 0$ that depends only on ν, ν', and c_1. Therefore, using the argument like that in the proof of Theorem 2.2 with an application of Lemma J.1 instead of Lemma 2.1, we obtain

$$\begin{aligned} \rho \leqslant C\big[&n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3 + (\psi^2 + \psi\beta)\bar{\varphi}(u) \\ &+\psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} + (\beta^{-1}\log p + \psi^{-1} + p^{-c})\sqrt{1 \vee \log(p\psi)}\big] \end{aligned} \tag{31}$$

where all notation is taken from the proof of Theorem 2.2. Recall that $\bar{\varphi}(\cdot)$ is any function satisfying $\bar{\varphi}(u) \geqslant \varphi(u)$ for all $u > 0$ and $\varphi(u) = \varphi_x(u) \vee \varphi_y(u)$. To bound $\varphi_x(u)$, we have

$$\begin{aligned} \bar{\mathrm{E}}[x_{ij}^2 1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\}] &\leqslant \bar{\mathrm{E}}[x_{ij}^4]/(u^2\bar{\mathrm{E}}[x_{ij}^2]) \\ &= \bar{\mathrm{E}}[x_{ij}^4]/(u^2\bar{\mathrm{E}}[x_{ij}^2]^2)\bar{\mathrm{E}}[x_{ij}^2] \leqslant (M_{4,2}^4/u^2)\bar{\mathrm{E}}[x_{ij}^2]. \end{aligned}$$

This implies that $\varphi_x(u) \leqslant M_{4,2}^2/u$. To bound $\varphi_y(u)$, note that $\bar{\mathrm{E}}[y_{ij}^4] \leqslant 3\bar{\mathrm{E}}[x_{ij}^4]$, which was shown in the proof of Theorem 2.2. In addition, $\bar{\mathrm{E}}[y_{ij}^2] = \bar{\mathrm{E}}[x_{ij}^2]$. Therefore, $\varphi_y(u) \leqslant C\varphi_x(u)$. Hence, we can set $\bar{\varphi}(u) := CM_{4,2}^2/u$ for all $u > 0$. The rest of the proof is the same as that for Theorem 2.2 with M_4 replaced by $M_{4,2}$. ∎

Proof of Corollary J.1. Note that in both cases, $M_{4,2}^2 \leqslant CB_n$ and

$$M_3^3 = \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|^3] \leqslant M_{3,2}^3 \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[x_{ij}^2]^{3/2} \leqslant CM_{3,2}^3 \leqslant CB_n.$$

Therefore, the claim of the corollary follows from Theorem J.1 by the same argument as that leading to Corollary 2.1 from Theorem 2.2. ∎

Proof of Theorem J.2. The proof is the same as that for Theorem 3.2 with Lemmas J.1 and J.2 replacing Lemmas 2.1 and 3.1. ∎

Proof of Corollary J.2. Since $B_n \geqslant 1$, both under (E.5) and under (E.6) we have $(\log(pn))^7/n \leqslant C_2 n^{-c_2}$. Let $\tilde{\zeta}_1 := \zeta_1 \vee n^{-1}$. Then conditions (14) and (15) hold with $(\tilde{\zeta}_1, \zeta_2)$ replacing (ζ_1, ζ_2) and $\tilde{\zeta}_1\sqrt{\log p} + \zeta_2 \leqslant Cn^{-c}$. Further, since $p \geqslant C_3 n^{c_3}$, we have $p^{-c}(1 \vee \log(p/\tilde{\zeta}_1))^{1/2} \leqslant Cn^{-c}$.

Let $\vartheta = \vartheta_n := ((\mathrm{E}[\Delta])^{1/2}/\log p) \vee n^{-1}$. Then $p^{-c}(1 \vee \log(p/\vartheta))^{1/2} \leqslant Cn^{-c}$. In addition, if $\vartheta = n^{-1}$, then $\vartheta^{1/3}(\log(p/\vartheta))^{2/3} \leqslant Cn^{-c}$. Finally,

$$M_4^2 = \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[x_{ij}^4]^{1/2} \leqslant M_{4,2}^2 \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[x_{ij}^2] \leqslant CM_{4,2}^2 \leqslant CB_n.$$

The rest of the proof is similar to that for Corollary 3.1. ∎

PROOF OF LEMMA J.1. In the proof, several constants will be introduced. All of these constants are implicitly assumed to depend only on c_1 and $\bar{\sigma}$.

We choose $c > 0$ such that $4\bar{\sigma}\sqrt{c} = c_1$. Fix $\varsigma > 0$. It suffices to consider the case $\varsigma \geqslant \bar{\sigma}p^{-c}$. Let $\underline{\sigma} := c_2 b_p/\sqrt{\log p}$ for sufficiently small $c_2 > 0$ to be chosen below. Note that $\underline{\sigma} \geqslant c_1 c_2$. So, if $\varsigma > \underline{\sigma}$, (30) holds trivially by selecting sufficiently large C.

Consider the case $\varsigma \leqslant \underline{\sigma}$. Assume that $z > b_p + \varsigma + \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}$. Then

$$\begin{aligned}\mathrm{P}(|\max_{1\leqslant j\leqslant p} Y_j - z| \leqslant \varsigma) &\leqslant \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \geqslant z - \varsigma)\\ &\leqslant \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \geqslant b_p + \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}) \leqslant \varsigma/\underline{\sigma}\end{aligned}$$

where the last inequality follows from Borell's inequality. So, (30) holds by selecting sufficiently large C.

Now assume that $z < b_p - \varsigma - \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}$. Then

$$\begin{aligned}\mathrm{P}(|\max_{1\leqslant j\leqslant p} Y_j - z| \leqslant \varsigma) &\leqslant \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant z + \varsigma)\\ &\leqslant \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant b_p - \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}) \leqslant \varsigma/\underline{\sigma}\end{aligned}$$

where the last inequality follows from Borell's inequality. So, (30) holds by selecting sufficiently large C.

Finally, assume that

$$b_p - \varsigma - \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)} \leqslant z \leqslant b_p + \varsigma + \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}.$$

Then

$$\mathrm{P}(|\max_{1\leqslant j\leqslant p} Y_j - z| \leqslant \varsigma) \leqslant \mathrm{P}(|\max_{j\in\tilde{J}} Y_j - z| \leqslant \varsigma) + \mathrm{P}(|\max_{j\in J\setminus\tilde{J}} Y_j - z| \leqslant \varsigma) =: I + II$$

where $J := \{1,\ldots,p\}$ and $\tilde{J} := \{j \in J : \sigma_j \leqslant \underline{\sigma}\}$. Consider I. We have

$$b_p \geqslant_{(1)} c_1\sqrt{\log p} =_{(2)} 4\bar{\sigma}\sqrt{c\log p} = 4\bar{\sigma}\sqrt{\log(p^c)} \geqslant_{(3)} 4\bar{\sigma}\sqrt{\log(\bar{\sigma}/\varsigma)}$$

where (1) holds by assumption, (2) follows from the definition of c, and (3) holds because $\varsigma \geqslant \bar{\sigma}p^{-c}$. In addition, there exists $C_2 > 0$ such that $\mathrm{E}[\max_{j\in\tilde{J}} Y_j] \leqslant c_2 C_2 b_p$, and there exist $C_3 > 0$ such that $b_p \leqslant C_3\bar{\sigma}\sqrt{\log p}$, so that $\underline{\sigma} \leqslant c_2 C_3 \bar{\sigma}$. We choose c_2 so that $c_2 C_2 \leqslant 1/4$, $c_2/\sqrt{\log p} \leqslant 1/8$, and $c_2 C_3 \leqslant 1$. Then $\underline{\sigma} \leqslant \bar{\sigma}$, $\underline{\sigma} \leqslant b_p/8$, and $\mathrm{E}[\max_{j\in\tilde{J}} Y_j] \leqslant b_p/4$. Also recall that $\varsigma \leqslant \underline{\sigma}$. Therefore,

$$\begin{aligned}b_p - 2\varsigma - \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)} - \mathrm{E}[\max_{j\in\tilde{J}} Y_j] &\geqslant b_p/2 - \bar{\sigma}\sqrt{2\log(\bar{\sigma}/\varsigma)}\\ &\geqslant \bar{\sigma}\sqrt{2\log(\bar{\sigma}/\varsigma)} \geqslant \underline{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}.\end{aligned}$$

So, Borell's inequality yields

$$I \leqslant \mathrm{P}(\max_{j\in\tilde{J}} Y_j \geqslant z-\varsigma) \leqslant \mathrm{P}(\max_{j\in\tilde{J}} Y_j \geqslant b_p - 2\varsigma - \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)}) \leqslant \varsigma/\underline{\sigma}$$

because $\sigma_j \leqslant \underline{\sigma}$ for all $j \in \tilde{J}$.

Consider II. It is proved in [11] that

$$II \leqslant 4\varsigma\{(1/\underline{\sigma} - 1/\bar{\sigma})|z| + a_p + 1\}/\underline{\sigma} \tag{32}$$

where $a_p := \mathrm{E}[\max_{j\in J\setminus\tilde{J}} Y_j/\sigma_j]$. See, in particular, equation (16) in that paper. Note that $a_p \leqslant b_p/\underline{\sigma}$. Therefore, (32) combined with our restriction on z yields

$$\begin{aligned} II &\leqslant 4\varsigma\left(2b_p + \varsigma + \bar{\sigma}\sqrt{2\log(\underline{\sigma}/\varsigma)} + \underline{\sigma}\right)/\underline{\sigma}^2 \\ &\leqslant 4\varsigma\left(2b_p + 2\underline{\sigma} + \bar{\sigma}\sqrt{2\log(\bar{\sigma}/\varsigma)}\right)/\underline{\sigma}^2 \end{aligned}$$

where in the second line we used the facts that $\varsigma \leqslant \underline{\sigma}$ and $\underline{\sigma} \leqslant \bar{\sigma}$ by assumption and by construction, respectively. Now (30) holds by selecting sufficiently large $C > 0$, and using the fact that $\underline{\sigma} \geqslant c_1c_2 > 0$. This completes the proof. ∎

Proof of Lemma J.2. The proof is the same as that for Theorem 2 in [11] with Lemma J.1 replacing Lemma 2.1. ∎

APPENDIX K: VALIDITY OF EFRON'S EMPIRICAL BOOTSTRAP

In this section, we study the validity of the empirical (or Efron's) bootstrap in approximating the distribution of T_0 in the simple case where x_{ij}'s are uniformly bounded (the bound can increase with n). Moreover, we consider here the asymptotics where $n \to \infty$ and possibly $p = p_n \to \infty$. Recall the setup in Section 2: let $x_1, \ldots, x_n$ be independent centered random vectors in $\mathbb{R}^p$ and define

$$T_0 = \max_{1\leqslant j\leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij}.$$

The empirical bootstrap procedure is described as follows. Let $x_1^*, \ldots, x_n^*$ be i.i.d. draws from the empirical distribution of $x_1, \ldots, x_n$. Conditional on $(x_i)_{i=1}^n$, $x_1^*, \ldots, x_n^*$ are i.i.d. with mean $\mathbb{E}_n[x_i]$ and covariance matrix $\mathbb{E}_n[(x_i - \mathbb{E}_n[x_i])(x_i - \mathbb{E}_n[x_i])']$. Define

$$T_0^* = \max_{1\leqslant j\leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n (x_{ij}^* - \mathbb{E}_n[x_{ij}]).$$

The empirical bootstrap approximates the distribution of T_0 by the conditional distribution T_0^* given $(x_i)_{i=1}^n$.

Recall

$$W_0 = \max_{1 \leqslant j \leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i x_{ij},$$

where $e_1, \ldots, e_n$ are i.i.d. $N(0,1)$ random variables independent of $(x_i)_{i=1}^n$. We shall here compare the conditional distribution of T_0^* to that of W_0.

THEOREM K.1. *Suppose that there exists constants $C_1 > c_1 > 0$ and a sequence $B_n \geqslant 1$ of constants such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$ and $|x_{ij}| \leqslant B_n$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$. Then provided that $B_n^2(\log(pn))^7 = o(n)$, with probability $1 - o(1)$,*

$$\sup_{t \in \mathbb{R}} |\mathrm{P}\{T_0^* \leqslant t \mid (x_i)_{i=1}^n\} - \mathrm{P}\{W_0 \leqslant t \mid (x_i)_{i=1}^n\}| = o(1).$$

This theorem shows the asymptotic equivalence of the empirical and Gaussian multiplier bootstraps. The validity of the empirical bootstrap (in the form similar to that in Theorem 3.1) follows relatively directly from the validity of the Gaussian multiplier bootstrap.

PROOF OF THEOREM K.1. The proof consists of three steps.

Step 1. We first show that with probability $1 - o(1)$, $c_1/2 \leqslant \mathbb{E}_n[(x_{ij} - \mathbb{E}_n[x_{ij}])^2] \leqslant 2C_1$ for all $1 \leqslant j \leqslant p$. By Lemma A.1,

$$\mathrm{E}\left[\max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[x_{ij}]|\right] \lesssim \sqrt{C_1(\log p)/n} + B_n(\log p)/n = o((\log p)^{-1/2}),$$

$$\mathrm{E}\left[\max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[x_{ij}^2] - \bar{\mathrm{E}}[x_{ij}^2]|\right] \lesssim \sqrt{C_1 B_n^2(\log p)/n} + B_n^2(\log p)/n = o(1),$$

so that uniformly in $1 \leqslant j \leqslant p$, $|\mathbb{E}_n[(x_{ij} - \mathbb{E}_n[x_{ij}])^2] - \bar{\mathrm{E}}[x_{ij}^2]| = o_{\mathrm{P}}(1)$, which implies the desired assertion.

Step 2. Define

$$\check{W}_0 = \max_{1 \leqslant j \leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i (x_{ij} - \mathbb{E}_n[x_{ij}]).$$

We show that with probability $1 - o(1)$,

$$\sup_{t \in \mathbb{R}} |\mathrm{P}\{T_0^* \leqslant t \mid (x_i)_{i=1}^n\} - \mathrm{P}\{\check{W}_0 \leqslant t \mid (x_i)_{i=1}^n\}| = o(1). \tag{33}$$

Conditional on $(x_i)_{i=1}^n$, $x_i^* - \mathbb{E}_n[x_i]$ are independent centered random vector in $\mathbb{R}^p$ with covariance matrix $\mathbb{E}_n[(x_i - \mathbb{E}_n[x_i])(x_i - \mathbb{E}_n[x_i])']$. Hence conditional on $(x_i)_{i=1}^n$, we can apply Corollary 2.1 to T_0^* to deduce (33).

Step 3. We show that with probability $1 - o(1)$,

$$\sup_{t \in \mathbb{R}} |\mathrm{P}\{\check{W}_0 \leqslant t \mid (x_i)_{i=1}^n\} - \mathrm{P}\{W_0 \leqslant t \mid (x_i)_{i=1}^n\}| = o(1).$$

By definition, we have

$$|\check{W}_0 - W_0| \leqslant \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[x_{ij}]| \times \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i \right| = o_{\mathrm{P}}((\log p)^{-1/2}).$$

Hence using the anti-concentration inequality together with Step 1, we deduce the desired assertion.

The conclusion of Theorem K.1 follows from combining Steps 1-3. ∎

APPENDIX L: COMPARISON OF OUR GAUSSIAN APPROXIMATION RESULTS TO OTHER ONES

We first point out that our Gaussian approximation result (4) can be viewed as a version of multivariate central limit theorem, which is concerned with conditions under which

$$|\mathrm{P}(X \in A) - \mathrm{P}(Y \in A)| \to 0, \tag{34}$$

uniformly in a collection of sets A, typically *all* convex sets. Recall that

$$X = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i,$$

where $x_1, \ldots, x_n$ are independent centered random vectors in $\mathbb{R}^p$ with possibly $p = p_n \to \infty$, and

$$Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i,$$

$y_1, \ldots, y_n$ independent random vectors with $y_i \sim N(0, \mathrm{E}[x_i x_i'])$. In fact, the result (4) in the main text can be rewritten as

$$\sup_{t \in \mathbb{R}} |\mathrm{P}\{X \in A_{\max}(t)\} - \mathrm{P}\{Y \in A_{\max}(t)\}| \to 0, \tag{35}$$

where $A_{\max}(t) = \{a \in \mathbb{R}^p : \max_{1 \leqslant j \leqslant p} a_j \leqslant t\}$.

Hence, our paper contributes to the literature on multivariate central limit theorems with growing number of dimensions (see, among others, [21, 24, 1, 15, 4]). These papers are concerned with results of the form (34), but either explicitly or implicitly require the condition that $p^c/n \to 0$ for some $c > 0$ (when specialized to a setting like our setup). Results in these papers

rely on the anti-concentration results for Gaussian random vectors on the δ-expansions of boundaries of arbitrary convex sets A (see [2]). We restrict our attention to the class of sets of the form $A_{\max}(t)$ in (35). These sets have a special structure that allows us to deal with the case where $p \gg n$: in particular, concentration of measure on the δ-expansion of boundary of $A_{\max}(t)$ is at most of order

$$\delta \mathrm{E}[\max_{j \leqslant p} Y_j]$$

for Gaussian random vectors with unit variance (and separable Gaussian processes more generally), as shown in [11] (see also Lemma 2.1).

There is large literature on bounding the difference:

$$|\mathrm{E}[H(X)] - \mathrm{E}[H(Y)]|, \tag{36}$$

for various smooth functions $H(\cdot) : \mathbb{R}^p \to \mathbb{R}$, in particular the recent work includes [14, 8, 25, 9]. Any such bounds lead to Gaussian approximations, though the structure of H's plays an important role in limiting the scope of this approximation. Two methods in the literature that turned out most fruitful for deriving gaussian approximation results in high dimensional settings ($p \to \infty$ as $n \to \infty$ in our context) are those of Lindeberg and Stein. The history of the Lindeberg method dates back to Lindeberg's original proof of the central limit theorem ([20]), which has been revived in the recent literature. We refer to the introduction of [7] for a brief history on the Lindeberg's method; see also [6]. The recent development on Stein's method when x_i's are multivariate can be found in [14, 8, 25, 9]. See also [5] for a comprehensive overview of different methods. In contrast to these papers, our paper analyzes a rather particular, yet important case $H(\cdot) = g(F_\beta(\cdot))$, with $g : \mathbb{R} \to \mathbb{R}$ (progressively less) smooth function and $0 < \beta \to \infty$, and in our case, self-normalized truncation, some fine properties of the smooth potential F_β, maximal fourth order moments of variables and of their envelops, play a critical role (as we comment further below), and so our main results can not be (immediately) deduced from these prior results (nor do we attempt to follow this route).

Using the Lindeberg's method and the smoothing technique of Bentkus [3], [22] derived in their Theorem 5 a Gaussian approximation result that is of similar form to our *simple* (non-main) Gaussian approximation result presented in Section I of the SM. However, in [22], the anti-concentration property is *assumed* (see equation (1.4) in their paper); in our notation, their assumption on anti-concentration says that there exists a constant C independent of n and p such that

$$\mathrm{P}(r < \max_{1 \leqslant j \leqslant p} |Y_j| \leqslant r + \epsilon) \leqslant C\epsilon(1 + r)^{-3}.$$

This assumption is useful for the analysis of the Donsker case, but does not apply in our (non-Donsker) cases. In fact, it rules out the simple case where $Y_1, \ldots, Y_p$ are independent (i.e., the coordinates in x_i are uncorrelated) or $Y_1, \ldots, Y_p$ are weakly dependent (i.e., the coordinates in x_i are weakly correlated). In addition, it is worth pointing out that the use of Bentkus's [3] smoothing[1] instead of the smoothing by potentials from spin glasses used here, does not lead to optimal results in our case, since very subtle properties (stability property noted in Lemma A.6) of potentials play an important role in our proofs, and in particular, is crucial for getting a reasonable exponent in the dependence on $\log p$.

Chatterjee [6], who also used the Lindeberg method, analyzed a spin-glass example like our example (E.4) and also derived a result similar to our *simple* (non-main) Gaussian approximation result presented in Section I of the SM, where $x_i = z_{ij}\epsilon_i$, where $(z_{ij})_{i=1}^p$ are fixed and $(\epsilon_i)_{i=1}^n$ are i.i.d $\mathbb{R}$-valued and centered with bounded third moment. We note [6] only provided a result for smooth functionals, but the extension to non-smooth cases follows from our Lemma 2.1 along with standard kernel smoothing. In fact, all of our paper is inspired by Chatterjee's work, and an early version employed (combinatorial) Lindeberg's method. We discuss the limitations of Lindeberg's approach (in the canonical form) in the main text, where we motivate the use of a combination of Slepian-Stein method in conjunction with self-normalized truncation and subtle properties of the potential function that approximates the maximum function. Generalization of results of Chatterjee to the case where $\mathbb{R}$-valued $(\epsilon_i)_{i=1}^n$ are locally dependent are given in [27], who uses Slepian-Stein methods for proofs and gave a result similar to our *simple* (non-main) Gaussian approximation result presented in Section I of the SM. Like [27], we also use a version of Slepian-Stein methods, but the proof details (as well as results and applications) are quite different, since we instead analyze the case where the data $(x_i)_{i=1}^n$ are $\mathbb{R}^p$-valued (instead of $\mathbb{R}$-valued) independent vectors, and since we have to perform truncation (to get good dependencies on the size of the envelopes, $\max_{1\leqslant j\leqslant p}|x_{ij}|$) and use subtle properties of the potential function to get our main results (to get good dependencies on $\log p$).

Using an interesting modification of the Lindeberg method, [17] obtained an invariance principle for a sequence of sub-Gaussian $\mathbb{R}$-valued random variables $(\epsilon_i)_{i=1}^n$ (instead of $\mathbb{R}^p$-valued case consider here). Specifically, they looked at the large-sample probability of $(\epsilon_i)_{i=1}^n$ hitting a polytope formed

[1] Bentkus gave a proof of existence of a smooth function that approximates a supremum norm, with certain properties. The use of these properties alone does not lead to sharp results in our case. We rely instead on smoothing by potentials from spin glasses, and their detailed properties, most importantly the stability property, noted in Lemma A.6, which is readily established for this smoothing method.

by p half-spaces, which is on the whole a different problem than studied here (though tools are insightful, e.g. the novel use of results developed by Nazarov [23]). These results have no intersection with our results, except for a special case of "sub-exponential regression/spin-glass example" (E.3), if we further require in that example, that $\varepsilon_{ij} = \varepsilon_i$ for all $1 \leqslant j \leqslant p$, that ε_i's are sub-Gaussian, that $\mathrm{E}[\varepsilon_i^3] = 0$, and that $B_n^2(\log p)^{16}/n \to 0$. All of these conditions and especially the last one are substantively more restrictive than what is obtained for the example (E.3) in our Corollary 2.1.

Finally, we note that when x_i's are identically distributed in addition to being independent, the theory of strong approximations and, in particular, Hungarian coupling can also be used to obtain results like that in (4) in the main text under conditions permitting $p \gg n$; see, for example, Theorem 3.1 in Koltchinskii [19] and Rio [26]. However, in order for this theory to work, x_i have to be well approximable in a Haar basis when considered as functions on the underlying probability space – e.g., $x_{ij} = f_{j,n}(u_i)$, where $f_{j,n}$ should have a total variation norm with respect to $u_i \sim U(0,1)^d$ (where d is fixed) that does not grow too quickly to enable the expansion in the Haar basis. This technique has been proven fruitful in many applications, but this requires a radically different structure than what our leading applications impose; instead our results, based upon Slepian-Stein methods, are more readily applicable in these settings (instead of controlling total variation bounds, they rely on control of maxima moments and moments of envelopes of $\{x_{ij}, j \leqslant p\}$). For further theoretical comparisons of the two methods in the context of strong approximations of suprema of non-Donsker emprical processes by those of Gaussian processes, in the classical kernel and series smoothing examples, we refer to our companion work [10] (there is no winner in terms of guaranteed rates of approximation, though side conditions seem to be weaker for the Slepian-Stein type methods; in particular Hungarian couplings often impose the boundedness conditions, e.g. $\|x_i\|_\infty \leqslant B_n$).

Supplementary Material III

Additional Application

APPENDIX M: 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 $v_i \in \mathbb{R}^d$ is a vector of non-stochastic covariates. The null hypothesis, H_0, is that there exists $\beta \in \mathbb{R}^d$ such that

$$\mathrm{E}[y_i] = v_i'\beta;\ i = 1, \dots, n. \tag{37}$$

The alternative hypothesis, H_a, is that there is no β satisfying (37). We allow for triangular array asymptotics so that everything in the model may depend on n. For brevity, however, we omit index n.

Let $\varepsilon_i = y_i - \mathrm{E}[y_i]$, $i = 1, \dots, n$. 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 $P_j(v_i)$, $j = 1, \dots, p$. 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 $j = 1, \dots, p$. 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 $\widehat{\varepsilon}_i = y_i - z_i'\widehat{\beta}$; $i = 1, \dots, n$ 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 a 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 $B_n \geqslant 1$ of constants such that for all $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, $1 \leqslant k \leqslant d$: (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 M.1 (Size Control of Adaptive Specification Test). *Let $c_2 > 0$ be some constant. Suppose that condition (S) is satisfied. Moreover, suppose that either*

(a) $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$ *for all* $1 \leqslant i \leqslant n$ *and* $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; or*
(b) $\mathrm{E}[\exp(|\varepsilon_i|/C_1)] \leqslant 2$ *for all* $1 \leqslant i \leqslant n$ *and* $B_n^2(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*.*

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 M.1. The literature on specification testing is large. In particular, [18] and [16] 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. ∎

Proof of Theorem M.1. We only consider case (a). The proof for case (b) is similar and hence omitted. In this proof, let c, c', C, C' denote generic positive constants depending only on $c_1, c_2, C_1, \underline{\sigma}^2, \bar{\sigma}^2$ and their values may change from place to place. Let

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

We make use of Corollary 3.1-(ii). To this end, we shall verify conditions (14) and (15) in Section 3 of the main text, which will be separately done in Steps 1 and 2, respectively.

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

By Corollary 2.1-(ii), we have

$$\begin{aligned}&\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}|>t\right)\\&\quad\leqslant \mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}|>t\right)+Cn^{-c},\end{aligned}$$

uniformly in $t\in\mathbb{R}$. By the Gaussian concentration inequality [Proposition A.2.1 30], for every $t>0$, we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}|>\mathrm{E}[\max_{1\leqslant j\leqslant p}|\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}|]+Ct\right)\leqslant e^{-t^2}.$$

Since $\mathrm{E}[\max_{1\leqslant j\leqslant p}|\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}|]\leqslant C\sqrt{\log p}$, we conclude that

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}|>C\sqrt{\log(pn)}\right)\leqslant C'n^{-c}. \tag{38}$$

Moreover,

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

Consider I_j. We have

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

where (1) follows from assumption S-(ii), (2) from S-(iv) and S-(v), and (3) from S-(vi). Since $\mathrm{E}[\|\mathbb{E}_n[v_i\varepsilon_i]\|^2]\leqslant C/n$, by Markov's inequality, for every $t>0$,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i-\varepsilon_i)^2]>t\right)\leqslant C/(nt). \tag{39}$$

Consider II_j. By Lemma A.1 and Markov's inequality, we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\varepsilon_i^2-\sigma_i^2)]|>t\right)\leqslant CB_n^2(\log p)/(\sqrt{n}t). \tag{40}$$

Consider III_j. We have $|III_j|\leqslant 2|\mathbb{E}_n[z_{ij}^2v_i'(\beta-\widehat{\beta})\varepsilon_i]|\leqslant 2\|\mathbb{E}_n[z_{ij}^2\varepsilon_i v_i]\|\|\widehat{\beta}-\beta\|$. Hence

$$\begin{aligned}&\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2\varepsilon_i(\widehat{\varepsilon}_i-\varepsilon_i)]|>t\right)\\&\quad\leqslant \mathrm{P}\left(\max_{1\leqslant j\leqslant p}\|\mathbb{E}_n[z_{ij}^2\varepsilon_i v_i]\|>t\right)+\mathrm{P}(\|\widehat{\beta}-\beta\|>1)\\&\quad\leqslant C[B_n^2(\log p)/(\sqrt{n}t)+1/n].\end{aligned} \tag{41}$$

By (39)-(41), we have

$$\text{(42)}\quad \mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|>t\right)\leqslant C[B_n^2(\log p)/(\sqrt{n}t)+1/(nt)+1/n].$$

In particular,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|>\underline{\sigma}^2/2\right)\leqslant Cn^{-c}.$$

Since $\mathbb{E}_n[z_{ij}^2\sigma_i^2]\geqslant\underline{\sigma}^2>0$ (which is guaranteed by S-(iii) and S-(ii)), on the event $\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|\leqslant\underline{\sigma}^2/2$, we have

$$\min_{1\leqslant j\leqslant p}\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]\geqslant\min_{1\leqslant j\leqslant p}\mathbb{E}_n[z_{ij}^2\sigma_i^2]-\underline{\sigma}^2/2\geqslant\underline{\sigma}^2/2,$$

and hence

$$\begin{aligned}|T-T_0|&=\max_{1\leqslant j\leqslant p}\left|\frac{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}-\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}\right|\times T_0\\&\leqslant C\max_{1\leqslant j\leqslant p}\left|\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}-\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}\right|\times T_0\\&\leqslant C\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2\sigma_i^2]-\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]|\times T_0,\end{aligned}$$

where the last step uses the simple fact that

$$|\sqrt{a}-\sqrt{b}|=\frac{|a-b|}{\sqrt{a}+\sqrt{b}}\leqslant\frac{|a-b|}{\sqrt{a}}.$$

By (38) and (42), for every $t>0$,

$$\mathrm{P}\left(|T-T_0|>Ct\sqrt{\log(pn)}\right)\leqslant C'[n^{-c}+B_n^2(\log p)/(\sqrt{n}t)+1/(nt)].$$

By choosing $t=(\log(pn))^{-1}n^{-c'}$ with 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 $0<t\leqslant\underline{\sigma}^2/2$, consider the event

$$\mathcal{E}=\left\{(\varepsilon_i)_{i=1}^n:\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|\leqslant t,\max_{1\leqslant i\leqslant p}(\widehat{\varepsilon}_i-\varepsilon_i)^2\leqslant t^2\right\}.$$

By calculations in Step 1, $\mathrm{P}(\mathcal{E}) \geqslant 1 - C[B_n^2(\log p)/(\sqrt{n}t) + 1/(nt^2) + 1/n]$. We shall show that, on this event,

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}| > C\sqrt{\log(pn)}\right) \leqslant n^{-1}, \tag{43}$$

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}(\widehat{\varepsilon}_i - \varepsilon_i) e_i/\sqrt{n}| > Ct\sqrt{\log(pn)}\right) \leqslant n^{-1}. \tag{44}$$

For (43), by the Gaussian concentration inequality, for every $s > 0$,

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}| > \mathrm{E}_e[\max_{1\leqslant j\leqslant p} |\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|] + Cs\right) \leqslant e^{-s^2}.$$

where we have used the fact $\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2] = \mathbb{E}_n[z_{ij}^2\sigma_i^2] + \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)] \leqslant \bar{\sigma}^2 + t \leqslant \bar{\sigma}^2 + \underline{\sigma}^2/2$ on the event $\mathcal{E}$. Here $\mathrm{E}_e[\cdot]$ means the expectation with respect to $(e_i)_{i=1}^n$ conditional on $(\varepsilon_i)_{i=1}^n$. Moreover, on the event $\mathcal{E}$,

$$\mathrm{E}_e[\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|] \leqslant C\sqrt{\log p}.$$

Hence by choosing $s = \sqrt{\log n}$, we obtain (43). Inequality (44) follows similarly, by noting that $(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} \leqslant \max_{1\leqslant i\leqslant n} |\widehat{\varepsilon}_i - \varepsilon_i| \leqslant t$ on the event $\mathcal{E}$.

Define

$$W_1 := \max_{1\leqslant j\leqslant p} \frac{|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}}.$$

Note that $\mathbb{E}_n[z_{ij}^2\sigma_i^2] \geqslant \underline{\sigma}^2$. Since on the event $\mathcal{E}$, $\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)]| \leqslant t \leqslant \underline{\sigma}^2/2$, in view of Step 1, on this event, we have

$$\begin{aligned}
|W - W_0| &\leqslant |W - W_1| + |W_1 - W_0| \\
&\leqslant CtW_1 + |W_1 - W_0| \\
&\leqslant Ct \max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}| + C \max_{1\leqslant j\leqslant p} |\sum_{i=1}^n z_{ij}(\widehat{\varepsilon}_i - \varepsilon_i) e_i/\sqrt{n}|.
\end{aligned}$$

Therefore, by (43) and (44), on the event $\mathcal{E}$, we have

$$\mathrm{P}_e\left(|W - W_0| > Ct\sqrt{\log(pn)}\right) \leqslant 2n^{-1}.$$

By choosing $t = (\log(pn))^{-1}n^{-c}$ with sufficiently small $c > 0$, we obtain the claim of this step.

Step 3. Steps 1 and 2 verified conditions (14) and (15) in Section 3 of the main text. Theorem M.1 case (a) follows from Corollary 3.1-(ii). ∎

REFERENCES

[1] Asriev, A.V. and Rotar, V.I. (1989). On the convergence rate in the infinite-dimensional central limit theorem for probabilities of hitting parallelepipeds. *Theory of Probability and Its Applications* **30** 691-701.

[2] Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* **10** 411-420.

[3] Bentkus, V. (1990). Smooth approximations of the norm and differentiable functions with bounded support in Banach space l_∞^k. *Lith. Math. Journal* **30** 223-230.

[4] Bentkus, V. (2003). On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Infer.* **113** 385-402.

[5] Bentkus, V., Götze, F., Paulauskas, V., and Rackauskus, L. (2000). The accuracy of Gaussian approximation in Banach spaces. In: *Limit Theorems of Probability Theory* (Eds. Y.V. Prokholov and V. Statulevičius), Springer, pp. 25-111.

[6] Chatterjee, S. (2005a). A simple invariance theorem. arXiv:math/0508213.

[7] Chatterjee, S. (2006). A generalization of Lindeberg's principle. *Ann. Probab.* **34** 2061-2076.

[8] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257-283.

[9] Chen, L., and Fang, X. (2011). Multivariate normal approximation by Stein's method: the concentration inequality approach. arXiv:1111.4073.

[10] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012a). Gaussian approximation of suprema of empirical processes. arXiv:1212.6906.

[11] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012b). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807. Submitted to *Probab. Theory Related Fields*.

[12] de la Peña, V., Lai, T. and Shao, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer.

[13] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

[14] Goldstein, L., and Rinott, Y. (1996) Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.* **33** 1-17.

[15] Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724-739.

[16] Guerre, E. and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *Ann. Statist.* **33** 840-870.

[17] Harsha, P., Klivans, A., and Meka, R. (2012). An invariance principle for polytopes. *Journal of the ACM* **59**.

[18] 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.

[19] 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.

[20] Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Math. Z.* **15** 211-225.

[21] 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.

[22] Norvaisa, R. and Paulauskas, V. (1991). Rate of convergence in the Central Limit Theorem for empirical processes. *Journal of Theoretical Probability.* **4** 511-534.

[23] Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to

a Gaussian measure. In: *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics Vol. 1807/2003, Springer, pp. 169-187.

[24] Portnoy, S. (1986). On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields* **73** 571-583.

[25] Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* **37** 2150-2173.

[26] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.

[27] Röllin, A. (2011). Stein's method in high dimensions with applications, *Ann. Inst. H. Poincaré Probab. Statist.* **49**, 529-549.

[28] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.

[29] Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians*. Springer.

[30] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Department of Economics and
Operations Research Center, MIT
50 Memorial Drive
Cambridge, MA 02142, USA.
E-mail: vchern@mit.edu

Department of Economics, UCLA
Bunche Hall, 8283
315 Portola Plaza
Los Angeles, CA 90095, USA.
E-mail: chetverikov@econ.ucla.edu

Graduate School of Economics
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-0033, Japan.
E-mail: kkato@e.u-tokyo.ac.jp