The Annals of Statistics
2013, Vol. 41, No. 6, 2786–2819
DOI: 10.1214/13-AOS1161

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

BY VICTOR CHERNOZHUKOV[1], DENIS CHETVERIKOV[1] AND KENGO KATO[2]

Massachusetts Institute of Technology, University of California, Los Angeles, 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 nonasymptotic 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

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

Received January 2013; revised June 2013.
[1]Supported by a National Science Foundation grant.
[2]Supported by the Grant-in-Aid for Young Scientists (B) (25780152), the Japan Society for the Promotion of Science.
MSC2010 subject classifications. 62E17, 62F40.
Key words 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:

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

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:

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

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

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

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, $|x_{ij}| \le C_1$ 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).

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

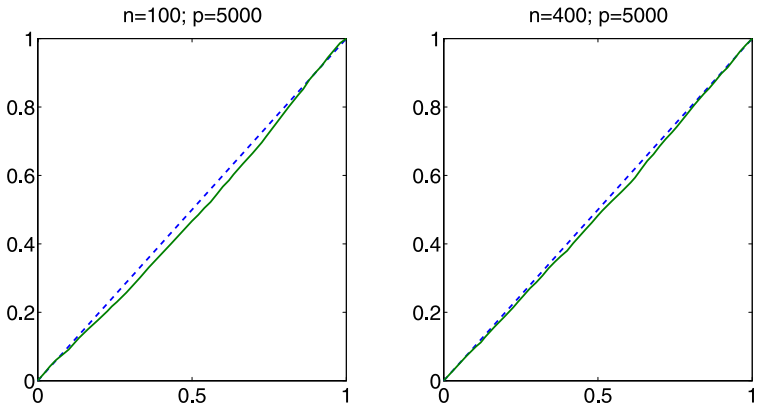

FIG. 1. *P–P plots comparing distributions of T_0 and Z_0 in the example motivated by the problem of selecting the penalty level of the Dantzig selector. Here x_{ij} are generated as $x_{ij} = z_{ij}\varepsilon_i$ with $\varepsilon_i \sim t(4)$, (a t-distribution with four degrees of freedom), and z_{ij} are nonstochastic (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.*

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, [11, 12, 19, 20, 33, 37, 39–41] 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 [15] 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 Appendix 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 $\{\sum_{i=1}^n x_{ij}/\sqrt{n}, 1 \le j \le p\}$

is not asymptotically Donsker (i.e., cannot 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 [20]. 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 $\mathrm{P}(Z_0 \le t)$ 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

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

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 $e_1, \dots, e_n$:

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

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:

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

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 [14], 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 [7, 29, 30, 36] 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., [21, 25] 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 $p \geq 3$. 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 $c, C, c_1, C_1, c_2, C_2, \ldots$ are understood to be independent of n. Throughout the paper, $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leq i \leq n$, 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, $A \ominus B = (A \setminus B) \cup (B \setminus A)$.

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:

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

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:

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

The problem of comparing distributions of maxima is of intrinsic difficulty since the maximum function $z = (z_1, \dots, z_p)' \mapsto \max_{1\le j\le p} z_j$ is nondifferentiable. To circumvent the problem, we use a smooth approximation of the maximum function. For $z = (z_1, \dots, z_p)' \in \mathbb{R}^p$, consider the function

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

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 $z \in \mathbb{R}^p$,

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

This smooth max function arises in the definition of "free energy" in spin glasses; see, for example, [41]. 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

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

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

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

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 $u \ge 0$ such that

$$\mathrm{P}\bigl(|x_{ij}| \le u\bigl(\bar{\mathrm{E}}\bigl[x_{ij}^2\bigr]\bigr)^{1/2}, 1 \le i \le n, 1 \le j \le p\bigr) \ge 1 - \gamma.$$

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:

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

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

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

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 $2\sqrt{2}uM_2\beta/\sqrt{n} \leq 1$ and $u \geq u(\gamma)$. 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*

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

where

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

We will also invoke the following lemma, which is proved in [15].

LEMMA 2.1 (Anti-concentration). (a) *Let $Y_1, \ldots, Y_p$ be jointly Gaussian random variables with $\mathrm{E}[Y_j] = 0$ and $\sigma_j^2 := \mathrm{E}[Y_j^2] > 0$ for all $1 \leq j \leq p$, and let $a_p := \mathrm{E}[\max_{1 \leq j \leq p}(Y_j/\sigma_j)]$. Let $\underline{\sigma} = \min_{1 \leq j \leq p} \sigma_j$ and $\bar{\sigma} = \max_{1 \leq j \leq p} \sigma_j$. Then for every $\varsigma > 0$,*

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

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-hand 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 $c_1 \leq \bar{\mathrm{E}}[x_{ij}^2] \leq C_1$ for all $1 \leq j \leq p$. Then for every $\gamma \in (0,1)$,*

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

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 $\bar{\mathrm{E}}[x_{ij}^2] \geq c_1$ for *all* $1 \leq j \leq p$ cannot 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 $J \subset \{1, \ldots, p\}$,

$$|J| \geq \nu p, \qquad \bar{\mathrm{E}}[x_{ij}^2] \geq c_1,$$

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

where $\nu > 0$ and $\nu' > 0$ are some constants independent of n or p. Appendix 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 [42], 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 $(\mathrm{E}[x_{ij}^2])^{1/2} \leq B$ for all $1 \leq i \leq n, 1 \leq j \leq p$ and $\bar{\mathrm{E}}[h(\max_{1 \leq j \leq p} |x_{ij}|/D)] \leq 1$. Then under the condition of Theorem 2.2,*

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

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 $B_n \geq 1$ 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* $1 \leq i \leq n$ and $1 \leq j \leq p$:

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

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

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

$$\bar{\mathrm{E}}\big[x_{ij}^2\big] \ge c_1 \quad \text{and} \quad \mathrm{E}\big[\exp\big(|x_{ij}|/C_1\big)\big] \le 2$$

(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 $|x_{ij}| \le B_n$ 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 $\max_{1\le j\le p} |x_{ij}|$.

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

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

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

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 $\mathrm{E}[\max_{1\le j\le p} \varepsilon_{ij}^4] \le C_1$ in condition (E.4) reduces to $\mathrm{E}[\varepsilon_i^4] \le C_1$. Interestingly, these examples are also connected to spin glasses, see, for example, [41] 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* $B_n^2(\log(pn))^7/n \le C_2 n^{-c_2}$ *or* (ii) (E.2) *or* (E.4) *holds and* $B_n^4(\log(pn))^7/n \le C_2 n^{-c_2}$. *Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, C_1, c_2, and C_2 such that*

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

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 Appendix 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* $c_1 \le \bar{\mathrm{E}}[x_{ij}^2] \le C_1$ *for all* $1 \le j \le p$. *Then there exists a constant* $C > 0$ *depending only on* c_1 *and* C_1 *such that*

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

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

This simple (though apparently new, at this level of generality) result follows from the classical Lindeberg's argument previously given in Chatterjee [10] [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, [14] 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 $\max_{1\le j\le p} V_j$ and $\max_{1\le j\le p} Y_j$ in terms of p and

$$\Delta_0 := \max_{1\le j,k\le p} |\Sigma_{jk}^V - \Sigma_{jk}^Y|.$$

LEMMA 3.1 (Comparison of distributions of Gaussian maxima). *Suppose that there are some constants* $0 < c_1 < C_1$ *such that* $c_1 \le \Sigma_{jj}^Y \le C_1$ *for all* $1 \le j \le p$. *Then there exists a constant* $C > 0$ *depending only on* c_1 *and* C_1 *such that*

$$\sup_{t\in\mathbb{R}} \Bigl|\mathrm{P}\Bigl(\max_{1\le j\le p} V_j \le t\Bigr) - \mathrm{P}\Bigl(\max_{1\le j\le p} Y_j \le t\Bigr)\Bigr| \le C\Delta_0^{1/3}\bigl(1 \vee \log(p/\Delta_0)\bigr)^{2/3}.$$

COMMENT 3.1. The result is derived in [15], 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

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

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

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

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,

$$c_{W_0}(\alpha) := \inf\bigl\{t \in \mathbb{R} : \mathrm{P}_e(W_0 \le t) \ge \alpha\bigr\},$$

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., $\mathrm{P}_e(W_0 \le t) = \mathrm{P}(W_0 \le t \mid (x_i)_{i=1}^n)$]. The multiplier bootstrap theorem below provides a nonasymptotic 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

$$c_{Z_0}(\alpha) := \inf\bigl\{t \in \mathbb{R} : \mathrm{P}(Z_0 \le t) \ge \alpha\bigr\},$$

where $Z_0 = \max_{1\le j\le p} \sum_{i=1}^n y_{ij}/\sqrt{n}$ and $(y_i)_{i=1}^n$ is a sequence of independent $N(0, \mathrm{E}[x_i x_i'])$ vectors. Recall that $\Delta = \max_{1\le j,k\le p} |\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]|$.

LEMMA 3.2 (Comparison of quantiles, I). *Suppose that there are some constants $0 < c_1 < C_1$ such that $c_1 \le \bar{\mathrm{E}}[x_{ij}^2] \le C_1$ for all $1 \le j \le p$. Then for every $\alpha \in (0,1)$,*

$$\mathrm{P}\bigl(c_{W_0}(\alpha) \le c_{Z_0}\bigl(\alpha + \pi(\vartheta)\bigr)\bigr) \ge 1 - \mathrm{P}(\Delta > \vartheta),$$
$$\mathrm{P}\bigl(c_{Z_0}(\alpha) \le c_{W_0}\bigl(\alpha + \pi(\vartheta)\bigr)\bigr) \ge 1 - \mathrm{P}(\Delta > \vartheta),$$

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

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

Recall that $\rho := \sup_{t\in\mathbb{R}} |\mathrm{P}(T_0 \le t) - \mathrm{P}(Z_0 \le t)|$. 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* $c_1 \le \bar{\mathrm{E}}[x_{ij}^2] \le C_1$ *for all* $1 \le j \le p$. *Then for every* $\vartheta > 0$,

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

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

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

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 $T = T(x_1, \ldots, x_n)$ 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 $W = W(x_1, \ldots, x_n, e_1, \ldots, e_n)$, 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 $\zeta_1 \ge 0$ and $\zeta_2 \ge 0$, depending on n (and typically $\zeta_1 \to 0$, $\zeta_2 \to 0$ as $n \to \infty$), such that

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

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

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

$$c_W(\alpha) := \inf\big\{t \in \mathbb{R} : \mathrm{P}_e(W \le t) \ge \alpha\big\}$$

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

$$\mathrm{P}\big(c_W(\alpha) \le c_{W_0}(\alpha + \zeta_2) + \zeta_1\big) \ge 1 - \zeta_2,$$
$$\mathrm{P}\big(c_{W_0}(\alpha) \le c_W(\alpha + \zeta_2) + \zeta_1\big) \ge 1 - \zeta_2.$$

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 $c_1 \le \bar{\mathrm{E}}[x_{ij}^2] \le C_1$ *for all* $1 \le j \le p$. *Moreover, suppose that* (14) *and* (15) *hold. Then for every* $\vartheta > 0$,

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

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

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 $T_0^* = \max_{1 \le j \le p} \sum_{i=1}^n (x_{ij}^* - \mathbb{E}_n[x_{ij}])/\sqrt{n}$ where $x_1, \ldots, x_n^*$ are i.i.d. draws from the empirical distribution of $x_1, \ldots, x_n$. We show in Appendix 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 $B_n \ge 1$ 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* $\zeta_1\sqrt{\log p} + \zeta_2 \le C_2 n^{-c_2}$. *Moreover, suppose that one of the following conditions is satisfied*: (i) (E.1) *or* (E.3) *holds and* $B_n^2(\log(pn))^7/n \le C_2 n^{-c_2}$ *or* (ii) (E.2) *or* (E.4) *holds and* $B_n^4(\log(pn))^7/n \le C_2 n^{-c_2}$. *Then there exist constants* $c > 0$ *and* $C > 0$ *depending only on* c_1, C_1, c_2 *and* C_2 *such that*

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

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

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 [6, 9, 43] in the Gaussian setting and of [28] 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 nonstochastic vector of regressors. We consider the model

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

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:

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

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 $\|\beta\|_0 \le s \ll n$ (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

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

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

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

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

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

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

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

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

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

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 $\mathrm{E}[\varepsilon_i^4] \le \text{const.}$ 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, $T_0 = \max_{1\le k\le 2p} n^{-1/2}\sum_{i=1}^n \tilde{z}_{ik}\varepsilon_i$, 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:

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

where $\mathcal{R}(\beta) := \{\delta\in\mathbb{R}^p : \|\beta+\delta\|_{\ell_1} \le \|\beta\|_{\ell_1}\}$ 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 (nonconic) 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 $\|\beta\|_0 \le s$, we have

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

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 $B_n \ge 1$ of constants such that for all $1 \le i \le n$ and $1 \le j \le p$*: (i) $|z_{ij}| \le B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\mathrm{E}[\varepsilon_i^2] = \sigma^2$; (iv) $\mathrm{E}[\varepsilon_i^4] \le C_1$; *and* (v) $B_n^4(\log(pn))^7/n \le C_1 n^{-c_1}$. *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,*

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

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

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

For example, in some designs, we can have $\sqrt{n}\max_{1\le j\le p}|\mathbb{E}_n[z_{ij}e_i]| = O_{\mathrm{P}}(1)$, 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 $\|\beta\|_0 \le s$. In this case, with probability at least $1-\alpha-Cn^{-c}$,

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

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

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

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

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

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

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

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* $B_n \ge 1$ *of constants such that for all* $1 \le i \le n$ *and* $1 \le j \le p$: (i) $|z_{ij}| \le B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\underline{\sigma}^2 \le \mathrm{E}[\varepsilon_i^2] \le \sigma^2$; (iv) $\mathrm{E}[\varepsilon_i^4] \le C_1$; (v) $B_n^4(\log(pn))^7/n \le C_1 n^{-c_1}$; *and* (vi) $(\log p)B_n c_0(1-1/n)/(\sqrt{n}\kappa_{\mathrm{pr}}(\beta)) \le C_1 n^{-c_1}$. *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*

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

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

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

where $c_{Z_0}(1-a) := (1-a)$*-quantile of* Z_0; *where* $c_{Z_0}(1-a) \le c_0(1-a)$.

COMMENT 4.5 (A portmanteu significance 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* $\widehat{\beta}^{(1)} \ne 0$, *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 $\|\widehat{\beta}^{(1)}\| \le \|\beta_0\| = 0$ 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* $\bigtimes_{j=1}^{p} I_j$, *where* $I_j = [\widehat{\beta}_j^{(1)} \pm 2\lambda^{(1)}/(\sqrt{n}\kappa_{\mathrm{jc}}(\beta))]$.

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 $1 \leq j \leq p$, 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 nonorthogonal 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:

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

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 $\mathcal{R}(\beta) := \{\delta \in \mathbb{R}^p : J(\beta+\delta) \leq J(\beta)\}$; 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 $H_j : \beta_j \le \beta_{0j}$ 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 $\mathcal{W} := \{1, \ldots, p\}$ and for $w \subset \mathcal{W}$ denote $\Omega^w := (\bigcap_{j \in w} \Omega_j) \cap (\bigcap_{j \notin w} \Omega_j^c)$ where $\Omega_j^c := \Omega \setminus \Omega_j$. The strong control of the family-wise error rate means

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

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

For $j = 1, \ldots, p$, 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 $l \ge 2$, 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

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

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

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,

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

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:

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

We will assume the following regularity condition:

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

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 $\underline{\sigma}^2 \le \mathrm{E}[x_{ij}^2] \le \bar{\sigma}^2$, $(\log p)^7 \le C_2 n^{1-c_2}$ for arbitrarily small c_2 and, for example, either (a) $\mathrm{E}[\exp(|x_{ij}|/C_1)] \le 2$ [condition (E.1)] or (b) $\mathrm{E}[\max_{1\le j\le p} x_{ij}^4] \le C_1$ [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 nonconservative.

APPENDIX A: PRELIMINARIES

A.1. A useful maximal inequality. The following lemma, which is derived in [15], 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 $p \ge 2$. Let $M = \max_{1\le i\le n} \max_{1\le j\le p} |x_{ij}|$ and $\sigma^2 = \max_{1\le j\le p} \bar{\mathrm{E}}[x_{ij}^2]$. Then*

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

PROOF. See [15], 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* $1 \le j,k,l \le p$,

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

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

$$\pi_j(z) := e^{\beta z_j} \Big/ \sum_{m=1}^p e^{\beta z_m}, \qquad w_{jk}(z) := (\pi_j\delta_{jk} - \pi_j\pi_k)(z),$$

$$q_{jkl}(z) := \big(\pi_j\delta_{jl}\delta_{jk} - \pi_j\pi_l\delta_{jk} - \pi_j\pi_k(\delta_{jl} + \delta_{kl}) + 2\pi_j\pi_k\pi_l\big)(z).$$

Moreover,

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

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

PROOF. The proof follows from the fact that $\partial_j F_\beta(z) = \pi_j(z)$ with $\pi_j(z) \ge 0$ 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* $1 \le j,k,l \le p$,

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

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. 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* $1 \le j,k,l \le p$,

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

where

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

Moreover,

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

PROOF. 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* $\max_{j\le p}|w_j|\beta \le 1$, $\tau \in [0,1]$, *and every* $1 \le j,k,l \le p$, *we have*

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

PROOF. Observe that

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

Similarly, $\pi_j(z+\tau w) \ge e^{-2}\pi_j(z)$. 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 [19], Theorem 2.16, for self-normalized sums.

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

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

PROOF. See Appendix 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

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

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

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

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

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

so that

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

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:

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

and

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

We have by Taylor's theorem

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

where

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

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

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

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

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

The claim of this step now follows.

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

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

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

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

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

Step 4 (Bound on *III*). Observe that

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

where (1) follows from $|\partial_j\partial_k\partial_l m(z)| \le U_{jkl}(z)$ (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:

right-hand side of (24)

$$\begin{aligned}
&\lesssim_{(4)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}\big[U_{jkl}(Z(t))\big]\cdot\mathrm{E}\big[\big|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)\big|\big]\,dt \\
&=_{(5)} \sum_{j,k,l=1}^{p}\int_0^1 \mathrm{E}\big[U_{jkl}(Z(t))\big]\cdot n\bar{\mathrm{E}}\big[\big|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)\big|\big]\,dt
\end{aligned}$$

$$\lesssim_{(6)} \int_0^1 \Bigg(\sum_{j,k,l=1}^{p} \mathrm{E}\big[U_{jkl}(Z(t))\big] \Bigg) \max_{1\le j,k,l\le p} n\bar{\mathrm{E}}\big[\big|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)\big|\big]\,dt$$

$$\lesssim_{(7)} \big(G_3 + G_2\beta + G_1\beta^2\big) \int_0^1 \max_{1\le j,k,l\le p} n\bar{\mathrm{E}}\big[\big|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)\big|\big]\,dt,$$

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

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

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

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

where (1) follows from the fact that: $|\dot{Z}_{ij}(t)/\omega(t)| \le (|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)/\sqrt{n}$, $|Z_{im}(t)| \le (|\tilde{x}_{im}| + |\tilde{y}_{im}|)/\sqrt{n}$, 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 $\max_{1\le j\le p} \bar{\mathrm{E}}[(|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3]$; (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 $\mathrm{E}[|y_{ij}|^3] \lesssim (\mathrm{E}[y_{ij}^2])^{3/2} = (\mathrm{E}[|x_{ij}^2|])^{3/2} \le \mathrm{E}[|x_{ij}|^3]$. 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 $\bar{\mathrm{E}}[x_{ij}^2] \ge c_1$ by assumption, we have $1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\} \le 1\{|x_{ij}| > c_1^{1/2}u\}$. By Markov's inequality and the con-

dition of the lemma, we have

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

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

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

Therefore, $u_y(\gamma) \le CB\sqrt{\log(pn/\gamma)}$ 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) $B \vee M_3^3 \vee M_4^2 \le CB_n$, $D \le CB_n \log p$, $h(v) = e^v - 1$;
(E.2) $B \vee D \vee M_3^3 \vee M_4^2 \le CB_n$, $h(v) = v^4$.

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 [42]. 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 $\ell_n := \log(pn^{1+c_3})$,

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

Hence taking $\gamma = n^{-c_3}$, we conclude from Theorem 2.2 and Lemma 2.2 that $\rho \le Cn^{-\min\{c_3,c_4\}}$ 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 $\Delta = \max_{1\le j,k\le p} |\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]|$. By Lemma 3.1, on the event $\{(x_i)_{i=1}^n : \Delta \le \vartheta\}$, we have $|\mathrm{P}(Z_0 \le t) - \mathrm{P}_e(W_0 \le t)| \le \pi(\vartheta)$ for all $t \in \mathbb{R}$, and so on this event

$$\begin{aligned}\mathrm{P}_e\big(W_0 \le c_{Z_0}(\alpha + \pi(\vartheta))\big) &\ge \mathrm{P}\big(Z_0 \le c_{Z_0}(\alpha + \pi(\vartheta))\big) - \pi(\vartheta)\\ &\ge \alpha + \pi(\vartheta) - \pi(\vartheta) = \alpha,\end{aligned}$$

implying the first claim. The second claim follows similarly.

C.2. Proof of Lemma 3.3. By equation (15), the probability of the event $\{(x_i)_{i=1}^n : \mathrm{P}_e(|W - W_0| > \zeta_1) \le \zeta_2\}$ is at least $1 - \zeta_2$. On this event,

$$\mathrm{P}_e\big(W \le c_{W_0}(\alpha + \zeta_2) + \zeta_1\big) \ge \mathrm{P}_e\big(W_0 \le c_{W_0}(\alpha + \zeta_2)\big) - \zeta_2 \ge \alpha + \zeta_2 - \zeta_2 = \alpha,$$

implying that $\mathrm{P}(c_W(\alpha) \le c_{W_0}(\alpha + \zeta_2) + \zeta_1) \ge 1 - \zeta_2$. 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

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

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

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

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*

$$\mathrm{E}[\Delta] \le C \times \begin{cases} \sqrt{\dfrac{B_n^2 \log p}{n}} \vee \dfrac{B_n^2 (\log(pn))^2 (\log p)}{n}, & \text{under (E.1)}, \\ \sqrt{\dfrac{B_n^2 \log p}{n}} \vee \dfrac{B_n^2 (\log p)}{\sqrt{n}}, & \text{under (E.2)}, \end{cases}$$

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

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

The conclusion of the lemma follows from elementary calculations with help of Lemma 2.2.2 in [42]. □

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, $\rho \le Cn^{-c}$. Moreover, $\zeta_1 \sqrt{\log p} \le C_2 n^{-c_2}$ implies that $\zeta_1 \le C_2 n^{-c_2}$ (recall $p \ge 3$), and hence $\zeta_1 \sqrt{\log(p/\zeta_1)} \le Cn^{-c}$. Also, $\zeta_2 \le Cn^{-c}$ by assumption.

Let $\vartheta = \vartheta_n := (\mathrm{E}[\Delta])^{1/2} / \log p$. By Lemma C.1, $\mathrm{E}[\Delta](\log p)^2 \le Cn^{-c}$. Therefore, $\pi(\vartheta) \le Cn^{-c}$ (with possibly different $c, C > 0$). In addition, by Markov's inequality, $\mathrm{P}(\Delta > \vartheta) \le \mathrm{E}[\Delta]/\vartheta \le Cn^{-c}$. Hence, by Theorem 3.2, the first inequality follows. The second inequality follows from the first inequality and the fact that $\rho \le Cn^{-c}$ 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" (DOI: 10.1214/13-AOS1161SUPP; .pdf). 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. MR2845886

[2] ANDERSON, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and early training projects. *J. Amer. Statist. Assoc.* **103** 1481–1495. MR2655725

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

[4] ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension. II. Multiple tests. *Ann. Statist.* **38** 83–99. MR2589317

[5] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

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

[7] BRETAGNOLLE, J. and MASSART, P. (1989). Hungarian constructions from the nonasymptotic viewpoint. *Ann. Probab.* **17** 239–256. MR0972783

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

[9] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[10] CHATTERJEE, S. (2005). A simple invariance theorem. Available at arXiv:math/0508213.

[11] CHATTERJEE, S. (2005). An error bound in the Sudakov–Fernique inequality. Available at arXiv:math/0510424.

[12] CHEN, L., GOLDSTEIN, L. and SHAO, Q. M. (2011). *Normal Approximation by Stein's Method*. Springer, New York.

[13] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2012). Central limit theorems and multiplier bootstrap when p is much larger than n. Available at arXiv:1212.6906.

[14] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2012). Gaussian approximation of suprema of empirical processes. Available at arXiv:1212.6906.

[15] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2012). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. Available at 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." DOI:10.1214/13-AOS1161SUPP.

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

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

[19] DE LA PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2009). *Self-Normalized Processes*: *Limit Theory and Statistical Applications*. Springer, Berlin. MR2488094

[20] DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge Univ. Press, Cambridge. MR1720712

[21] DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961

[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. Statist. Assoc.* **102** 1282–1288. MR2372536

[23] FRICK, K., MARNITZ, P. and MUNK, A. (2012). Shape-constrained regularization by statistical multiresolution for inverse problems: Asymptotic analysis. *Inverse Problems* **28** 065006, 31. MR2924303

[24] GAUTIER, E. and TSYBAKOV, A. (2011). High-dimensional istrumental variables regression and confidence sets. Available at arXiv:1105.2454.

[25] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707

[26] HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124

[27] JUDITSKY, A. and NEMIROVSKI, A. (2011). On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Math. Program.* **127** 57–88. MR2776710

[28] KOLTCHINSKII, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200

[29] KOLTCHINSKII, V. I. (1994). Komlos–Major–Tusnady approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab.* **7** 73–118. MR1256392

[30] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131. MR0375412

[31] LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York. MR0691492

[32] MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176

[33] PANCHENKO, D. (2013). *The Sherrington–Kirkpatrick Model*. Springer, New York. MR3052333

[34] POLLARD, D. (2002). *A User's Guide to Measure Theoretic Probability*. *Cambridge Series in Statistical and Probabilistic Mathematics* **8**. Cambridge Univ. Press, Cambridge. MR1873379

[35] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301

[36] RIO, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21–45. MR1254823

[37] RÖLLIN, A. (2011). Stein's method in high dimensions with applications. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 529–549.

[38] ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821

[39] SLEPIAN, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.* **41** 463–501. MR0133183

[40] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[41] TALAGRAND, M. (2003). *Spin Glasses*: *A Challenge for Mathematicians*: *Cavity and Mean Field Models*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge*. *A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas*. 3*rd Series*. *A Series of Modern Surveys in Mathematics*] **46**. Springer, Berlin. MR1993891

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

[43] YE, F. and ZHANG, C.-H. (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. MR2756192

V. CHERNOZHUKOV
DEPARTMENT OF ECONOMICS
AND OPERATIONS RESEARCH CENTER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02142
USA
E-MAIL: vchern@mit.edu

D. CHETVERIKOV
DEPARTMENT OF ECONOMICS
UNIVERSITY OF CALIFORNIA, LOS ANGELES
BUNCHE HALL, 8283
315 PORTOLA PLAZA
LOS ANGELES, CALIFORNIA 90095
USA
E-MAIL: chetverikov@econ.ucla.edu

K. KATO
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