The Annals of Statistics
2020, Vol. 48, No. 2, 1214–1229
https://doi.org/10.1214/19-AOS1844

BOOTSTRAPPING MAX STATISTICS IN HIGH DIMENSIONS: NEAR-PARAMETRIC RATES UNDER WEAK VARIANCE DECAY AND APPLICATION TO FUNCTIONAL AND MULTINOMIAL DATA

BY MILES E. LOPES[*], ZHENHUA LIN[**] AND HANS-GEORG MÜLLER[†]

Department of Statistics, University of California, Davis, [*]melopes@ucdavis.edu; [**]zhnlin@ucdavis.edu; [†]hgmueller@ucdavis.edu

In recent years, bootstrap methods have drawn attention for their ability to approximate the laws of "max statistics" in high-dimensional problems. A leading example of such a statistic is the coordinatewise maximum of a sample average of n random vectors in $\mathbb{R}^p$. Existing results for this statistic show that the bootstrap can work when $n \ll p$, and rates of approximation (in Kolmogorov distance) have been obtained with only logarithmic dependence in p. Nevertheless, one of the challenging aspects of this setting is that established rates tend to scale like $n^{-1/6}$ as a function of n.

The main purpose of this paper is to demonstrate that improvement in rate is possible when extra model structure is available. Specifically, we show that if the coordinatewise variances of the observations exhibit decay, then a nearly $n^{-1/2}$ rate can be achieved, *independent of* p. Furthermore, a surprising aspect of this dimension-free rate is that it holds even when the decay is *very weak*. Lastly, we provide examples showing how these ideas can be applied to inference problems dealing with functional and multinomial data.

1. Introduction. One of the current challenges in theoretical statistics is to understand when bootstrap methods work in high-dimensional problems. In this direction, there has been a surge of recent interest in connection with "max statistics" such as

$$T = \max_{1 \le j \le p} S_{n,j},$$

where $S_{n,j}$ is the jth coordinate of the sum $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$, involving i.i.d. vectors $X_1, \ldots, X_n$ in $\mathbb{R}^p$.

This type of statistic has been a focal point in the literature for at least two reasons. First, it is an example of a statistic for which bootstrap methods can succeed in high dimensions under mild assumptions, which was established in several pathbreaking works (Arlot, Blanchard and Roquain (2010a), Arlot, Blanchard and Roquain (2010b), Chernozhukov, Chetverikov and Kato (2013), Chernozhukov, Chetverikov and Kato (2017)). Second, the statistic T is closely linked to several fundamental topics, such as suprema of empirical processes, nonparametric confidence regions and multiple testing problems. Likewise, many applications of bootstrap methods for max statistics have ensued at a brisk pace in recent years (see, e.g., Belloni et al. (2018), Chang, Yao and Zhou (2017), Chen (2018), Chen, Genovese and Wasserman (2015), Chernozhukov, Chetverikov and Kato (2014), Dezeure, Bühlmann and Zhang (2017), Fan, Shao and Zhou (2018), Wasserman, Kolar and Rinaldo (2014), Zhang and Cheng (2017)).

One of the favorable aspects of bootstrap approximation results for the distribution $\mathcal{L}(T)$ is that rates have been established with only logarithmic dependence in p. For instance, the

Received July 2018; revised January 2019.
MSC2010 subject classifications. Primary 62G09, 62G15; secondary 62G05, 62G20.
Key words and phrases. Bootstrap, high-dimensional statistics, rate of convergence, functional data analysis, multinomial data, confidence region, hypothesis test.

results in Chernozhukov, Chetverikov and Kato (2017) imply that under certain conditions, the Kolmogorov distance d_{K} between $\mathcal{L}(T)$ and its bootstrap counterpart $\mathcal{L}(T^*|X)$ satisfies the bound

$$d_{\mathrm{K}}\big(\mathcal{L}(T), \mathcal{L}\big(T^*|X\big)\big) \leq \frac{c\log(p)^b}{n^{1/6}} \tag{1.1}$$

with high probability, where $c, b > 0$ are constants not depending on n or p, and X denotes the matrix whose rows are $X_1, \ldots, X_n$. (In the following, symbols such as c will be often reused to designate a positive constant not depending on n or p, possibly with a different value at each occurrence.) Additional refinements of this result can be found in the same work, with regard to the choice of metric, or choice of bootstrap method. Also, recent progress in sharpening the exponent b has been made by Deng and Zhang (2017). However, this mild dependence on p is offset by the $n^{-1/6}$ dependence on n, which differs from the $n^{-1/2}$ dependence in the multivariate Berry–Esseen theorem when $p \ll n$.

Currently, the general problem of determining the best possible rates for Gaussian and bootstrap approximations is largely open in the high-dimensional setting. In particular, if we let $\tilde{T}$ denote the counterpart of T that arises from replacing $X_1, \ldots, X_n$ with independent Gaussian vectors $\tilde{X}_1, \ldots, \tilde{X}_n$ satisfying $\operatorname{cov}(X_i) = \operatorname{cov}(\tilde{X}_i)$, then a conjecture of Chernozhukov, Chetverikov and Kato (2017) indicates that a bound of the form $d_{\mathrm{K}}(\mathcal{L}(T), \mathcal{L}(\tilde{T})) \leq cn^{-1/6}\log(p)^b$ is optimal under certain conditions. A related conjecture in the setting of high-dimensional U-statistics may also be found in Chen (2018). (Further discussion of related work on Gaussian approximation is given in Appendix H.) Nevertheless, the finite-sample performance of bootstrap methods for max statistics is often more encouraging than what might be expected from the $n^{-1/6}$ dependence on n (see, e.g., Belloni et al. (2018), Fan, Shao and Zhou (2018), Zhang and Cheng (2017)). This suggests that improved rates are possible in at least some situations.

The purpose of this paper is to quantify an instance of such improvement when additional model structure is available. Specifically, we consider the case when the coordinates of $X_1, \ldots, X_n$ have decaying variances. If we let $\sigma_j^2 = \operatorname{var}(X_{1,j})$ for each $1 \leq j \leq p$, and write $\sigma_{(1)} \geq \cdots \geq \sigma_{(p)}$, then this condition may be formalized as

$$\sigma_{(j)} \leq cj^{-\alpha} \quad \text{for all } j \in \{1, \ldots, p\}, \tag{1.2}$$

where $\alpha > 0$ is a parameter not depending on n or p. (A complete set of assumptions, including a weaker version of (1.2), is given in Section 2.) This type of condition arises in many contexts, and in Section 2 we discuss examples related to principal component analysis, count data and Fourier coefficients of functional data. Furthermore, this condition can be assessed in practice, due to the fact that the parameters $\sigma_1, \ldots, \sigma_p$ can be accurately estimated, even in high dimensions (cf. Lemma D.7).

Within the setting of decaying variances, our main results show that a nearly parametric rate can be achieved for both Gaussian and bootstrap approximation of $\mathcal{L}(T)$. More precisely, this means that for any fixed $\delta \in (0, 1/2)$, the bound $d_{\mathrm{K}}(\mathcal{L}(T), \mathcal{L}(\tilde{T})) \leq cn^{-1/2+\delta}$ holds, and similarly, the event

$$d_{\mathrm{K}}\big(\mathcal{L}(T), \mathcal{L}\big(T^*|X\big)\big) \leq cn^{-1/2+\delta} \tag{1.3}$$

holds with high probability. Here, it is worth emphasizing a few basic aspects of these bounds. First, they are nonasymptotic and *do not depend on* p. Second, the parameter α is allowed to be *arbitrarily small*, and in this sense, the decay condition (1.2) is very weak. Third, the result for T^* holds when it is constructed using the standard multiplier bootstrap procedure (Chernozhukov, Chetverikov and Kato (2013)).

With regard to the existing literature, it is important to clarify that our near-parametric rate does not conflict with the conjectured optimality of the rate $n^{-1/6}$ for Gaussian approximation. The reason is that the $n^{-1/6}$ rate has been established in settings where the values $\sigma_1, \ldots, \sigma_p$ are restricted from becoming too small. A basic version of such a requirement is that

$$\min_{1 \le j \le p} \sigma_j \ge c. \tag{1.4}$$

Hence, the conditions (1.2) and (1.4) are complementary. Also, it is interesting to observe that the two conditions "intersect" in the limit $\alpha \to 0^+$, suggesting there is a phase transition in rates at the "boundary" corresponding to $\alpha = 0$.

Another important consideration that is related to the conditions (1.2) and (1.4) is the use of standardized variables. Namely, it is of special interest to approximate the distribution of the statistic

$$T' = \max_{1 \le j \le p} S_{n,j}/\sigma_j,$$

which is equivalent to approximating $\mathcal{L}(T)$ when each $X_{i,j}$ is standardized to have variance 1. Given that standardization eliminates variance decay, it might seem that the rate $n^{-1/2+\delta}$ has no bearing on approximating $\mathcal{L}(T')$. However, it is still possible to take advantage of variance decay, by using a basic notion that we refer to as "partial standardization."

The idea of partial standardization is to slightly modify T' by using a fractional power of each σ_j. Specifically, if we let $\tau_n \in [0, 1]$ be a free parameter, then we can consider the partially standardized statistic

$$M = \max_{1 \le j \le p} S_{n,j}/\sigma_j^{\tau_n}, \tag{1.5}$$

which interpolates between T and T' as τ_n ranges over $[0, 1]$. This statistic has the following significant property: If $X_1, \ldots, X_n$ satisfy the variance decay condition (1.2), and if τ_n is chosen to be slightly less than 1, then our main results show that the rate $n^{-1/2+\delta}$ holds for bootstrap approximations of $\mathcal{L}(M)$. In fact, this effect occurs even when $\tau_n \to 1$ as $n \to \infty$. Further details can be found in Section 3. Also note that our main results are formulated entirely in terms of M, which covers the statistic T as a special case.

In practice, simultaneous confidence intervals derived from approximations to $\mathcal{L}(M)$ are just as easy to use as those based on $\mathcal{L}(T')$. Although there is a slight difference between the quantiles of M and T' when $\tau_n < 1$, the important point is that the quantiles of $\mathcal{L}(M)$ may be preferred, since faster rates of bootstrap approximation are available. (See also Figure 1 in Section 4.) In this way, the statistic M offers a simple way to blend the utility of standardized variables with the beneficial effects of variance decay.

Outline. The remainder of the paper is organized as follows. In Section 2, we outline the problem setting, with a complete statement of the theoretical assumptions, as well as some motivating facts and examples. Our main results are given in Section 3, which consist of a Gaussian approximation result for $\mathcal{L}(M)$ (Theorem 3.1), and a corresponding bootstrap approximation result (Theorem 3.2). To provide a numerical illustration of our results, we discuss a problem in functional data analysis in Section 4, where the variance decay condition naturally arises. Specifically, we show how bootstrap approximations to $\mathcal{L}(M)$ can be used to derive simultaneous confidence intervals for the Fourier coefficients of a mean function. A second application to high-dimensional multinomial models is described in Section 5, which offers both a theoretical bootstrap approximation result, as well as some numerical results. Lastly, our conclusions are summarized in Section 6. All proofs are given in the Appendices, found in the Supplementary Material (Lopes, Lin and Müller (2020)).

Notation. The standard basis vectors in $\mathbb{R}^p$ are denoted $e_1, \dots, e_p$, and the identity matrix of size $p \times p$ is denoted $\mathbf{I}_p$. For any symmetric matrix $A \in \mathbb{R}^{p \times p}$, the ordered eigenvalues are denoted $\lambda(A) = (\lambda_1(A), \dots, \lambda_p(A))$, where $\lambda_{\max}(A) = \lambda_1(A) \geq \cdots \geq \lambda_p(A) = \lambda_{\min}(A)$. The operator norm of a matrix A, denoted $\|A\|_{\mathrm{op}}$, is the same as its largest singular value. If $v \in \mathbb{R}^p$ is a fixed vector, and $r > 0$, we write $\|v\|_r = (\sum_{j=1}^p |v_j|^r)^{1/r}$. In addition, the weak-$\ell_r$ (quasi) norm is given by $\|v\|_{w\ell_r} = \max_{1 \leq j \leq p} j^{1/r} |v|_{(j)}$, where $|v|_{(1)} \geq \cdots \geq |v|_{(p)}$ are the sorted absolute entries of v. Likewise, the notation $v_{(1)} \geq \cdots \geq v_{(p)}$ refers to the sorted entries. In a slight abuse of notation, we write $\|\xi\|_r = \mathbb{E}[|\xi|^r]^{1/r}$ to refer to the L^r norm of a scalar random variable ξ, with $r \geq 1$. The ψ_1-Orlicz norm is $\|\xi\|_{\psi_1} = \inf\{t > 0 \mid \mathbb{E}[\exp(|\xi|/t)] \leq 2\}$. If $\{a_n\}$ and $\{b_n\}$ are sequences of nonnegative real numbers, then the relation $a_n \lesssim b_n$ means that there is a constant $c > 0$ not depending on n, and an integer $n_0 \geq 1$, such that $a_n \leq c b_n$ for all $n \geq n_0$. Also, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Lastly, define the abbreviations $a_n \vee b_n = \max\{a_n, b_n\}$ and $a_n \wedge b_n = \min\{a_n, b_n\}$.

2. Setting and preliminaries. We consider a sequence of models indexed by n, with all parameters depending on n, except for those that are stated to be fixed. In particular, the dimension $p = p(n)$ is regarded as a function of n, and hence, if a constant does not depend on n, then it does not depend on p either.

ASSUMPTION 2.1 (Data-generating model).

(i) There is a vector $\mu = \mu(n) \in \mathbb{R}^p$ and positive semidefinite matrix $\Sigma = \Sigma(n) \in \mathbb{R}^{p \times p}$, such that the observations $X_1, \dots, X_n \in \mathbb{R}^p$ are generated as $X_i = \mu + \Sigma^{1/2} Z_i$ for each $1 \leq i \leq n$, where the random vectors $Z_1, \dots, Z_n \in \mathbb{R}^p$ are i.i.d.

(ii) The random vector Z_1 satisfies $\mathbb{E}[Z_1] = 0$ and $\mathbb{E}[Z_1 Z_1^\top] = \mathbf{I}_p$, as well as $\sup_{\|u\|_2 = 1} \|Z_1^\top u\|_{\psi_1} \leq c_0$, for some constant $c_0 > 0$ that does not depend on n.

Remarks. Note that no constraints are placed on the ratio p/n. Also, the subexponential tail condition in part (ii) is similar to other tail conditions that have been used in previous works on bootstrap methods for max statistics (Chernozhukov, Chetverikov and Kato (2013), Deng and Zhang (2017)).

To state our next assumption, it is necessary to develop some notation. For any $d \in \{1, \dots, p\}$, let $\mathcal{J}(d)$ denote a set of indices corresponding to the d largest values among $\sigma_1, \dots, \sigma_p$, that is, $\{\sigma_{(1)}, \dots, \sigma_{(d)}\} = \{\sigma_j \mid j \in \mathcal{J}(d)\}$. In addition, let $R(d) \in \mathbb{R}^{d \times d}$ denote the correlation matrix of the random variables $\{X_{1,j} \mid j \in \mathcal{J}(d)\}$. Lastly, let $a \in (0, \frac{1}{2})$ be a constant fixed with respect to n, and define the integers ℓ_n and k_n according to

$$\ell_n = \lceil (1 \vee \log(n)^3) \wedge p \rceil, \tag{2.1}$$

$$k_n = \lceil (\ell_n \vee n^{\frac{1}{\log(n)^a}}) \wedge p \rceil. \tag{2.2}$$

Note that both ℓ_n and k_n grow slower than any fractional power of n, and always satisfy $1 \leq \ell_n \leq k_n \leq p$.

ASSUMPTION 2.2 (Structural assumptions).

(i) The parameters $\sigma_1, \dots, \sigma_p$ are positive, and there are positive constants α, c, and $c_\circ \in (0, 1)$, not depending on n, such that

$$\sigma_{(j)} \leq c j^{-\alpha} \quad \text{for all } j \in \{k_n, \dots, p\}, \tag{2.3}$$

$$\sigma_{(j)} \geq c_\circ j^{-\alpha} \quad \text{for all } j \in \{1, \dots, k_n\}. \tag{2.4}$$

(ii) There is a constant $\epsilon_0 \in (0, 1)$, not depending on n, such that

$$\max_{i \neq j} R_{i,j}(\ell_n) \leq 1 - \epsilon_0. \tag{2.5}$$

Also, the matrix $R^+(\ell_n)$ with (i, j) entry given by $\max\{R_{i,j}(\ell_n), 0\}$ is positive semidefinite, and there is a constant $C > 0$ not depending on n such that

$$\sum_{1 \leq i < j \leq \ell_n} R^+_{i,j}(\ell_n) \leq C\ell_n. \tag{2.6}$$

Remarks. Since $\ell_n, k_n \ll n$, it is possible to accurately estimate the parameters $\sigma_{(1)}, \ldots, \sigma_{(k_n)}$, as well as the matrix $R(\ell_n)$, even when p is large (cf. Lemmas D.6 and D.7). In this sense, it is possible to empirically assess the conditions above. When considering the size of the decay parameter α, note that if Σ is viewed as a covariance operator acting on a Hilbert space, then the condition $\alpha > 1/2$ essentially corresponds to the case of a trace-class operator—a property that is typically assumed in functional data analysis (Hsing and Eubank (2015)). From this perspective, the condition $\alpha > 0$ is very weak, and allows the trace of Σ to diverge as $p \to \infty$.

With regard to the conditions on the correlation matrix $R(\ell_n)$, it is important to keep in mind that they only apply to a small set of variables of size $\mathcal{O}(\log(n)^3)$, and the dependence among the variables outside of $\mathcal{J}(\ell_n)$ is *completely unrestricted*. The interpretation of (2.6) is that it prevents excessive dependence among the coordinates with the largest variances. Meanwhile, the condition that $R^+(\ell_n)$ is positive semidefinite is more technical in nature, and is only used in order to apply a specialized version of Slepian's lemma (Lemma G.3). Nevertheless, this condition always holds in the important case where $R(\ell_n)$ is nonnegative. Perturbation arguments may also be used to obtain other examples where some entries of $R(\ell_n)$ are negative.

2.1. *Examples of correlation matrices.* Some correlation matrices satisfying Assumption 2.2(ii) are given below.

- *Autoregressive*: $R_{i,j} = \rho_0^{|i-j|}$ for any $\rho_0 \in (0, 1)$.
- *Algebraic decay*: $R_{i,j} = 1\{i = j\} + \dfrac{1\{i \neq j\}}{4|i - j|^\gamma}$ for any $\gamma \geq 2$.
- *Banded*: $R_{i,j} = \left(1 - \dfrac{|i - j|}{c_0}\right)_+$ for any $c_0 > 0$.
- *Multinomial*: $R_{i,j} = 1\{i = j\} - \sqrt{\dfrac{\pi_i \pi_j}{(1 - \pi_i)(1 - \pi_j)}}\, 1\{i \neq j\}$,

  where $(\pi_1, \ldots, \pi_p)$ is a probability vector.

By combining these types of correlation matrices with choices of $(\sigma_1, \ldots, \sigma_p)$ that satisfy (2.3) and (2.4), it is straightforward to construct examples of Σ that satisfy all aspects of Assumption 2.2.

2.2. *Examples of variance decay.* To provide additional context for the decay condition (2.3), we describe some general situations where it occurs.

- *Principal component analysis (PCA).* The broad applicability of PCA rests on the fact that many types of data have an underlying covariance matrix with weakly sparse eigenvalues. Roughly speaking, this means that most of the eigenvalues of Σ are small in comparison

to the top few. Similar to the condition (2.3), this situation can be modeled with the decay condition

$$\lambda_j(\Sigma) \le cj^{-\gamma}, \tag{2.7}$$

for some parameter $\gamma > 0$ (e.g., Bunea and Xiao (2015)). Whenever this holds, it can be shown that the variance decay condition *must* hold for some associated parameter $\alpha > 0$, and this is done in Proposition 2.1 below. So, in a qualitative sense, this indicates that if a dataset is amenable to PCA, then it is also likely to fall within the scope of our setting.

Another way to see the relationship between PCA and variance decay is through the measure of "effective rank," defined as

$$\mathtt{r}(\Sigma) = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}. \tag{2.8}$$

This quantity has played a key role in a substantial amount of recent work on PCA, because it offers a useful way to describe covariance matrices with an "intermediate" degree of complexity, which may be neither very low dimensional, nor very high dimensional. We refer to Vershynin (2012), Lounici (2014), Bunea and Xiao (2015), Reiß and Wahl (2020), Koltchinskii and Lounici (2017a), Koltchinskii and Lounici (2017b), Koltchinskii, Löffler and Nickl (2019+), Naumov, Spokoiny and Ulyanov (2019), and Jung, Lee and Ahn (2018), among others. Many of these works have focused on regimes where

$$\mathtt{r}(\Sigma) = o(n), \tag{2.9}$$

which conforms naturally with variance decay. Indeed, within a basic setup where $n \asymp p$ and $\|\Sigma\|_{\mathrm{op}} \asymp 1$, the condition (2.9) holds under $\sigma_{(j)} \le cj^{-\alpha}$ for any $\alpha > 0$.

- *Count data.* Consider a multinomial model based on p cells and n trials, parameterized by a vector of cell proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$. If the ith trial is represented as a vector $X_i \in \mathbb{R}^p$ in the set of standard basis vectors $\{e_1, \dots, e_p\}$, then the marginal distributions of X_i are binomial with $\sigma_j^2 = \pi_j(1 - \pi_j)$. In particular, it follows that *all* multinomial models satisfy the variance decay condition (2.3), because if we let $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_p)$, then the weak-$\ell_2$ norm of $\boldsymbol{\sigma}$ must satisfy $\|\boldsymbol{\sigma}\|_{w\ell_2} \le \|\boldsymbol{\sigma}\|_2 \le 1$, which implies

$$\sigma_{(j)} \le j^{-1/2} \tag{2.10}$$

for all $j \in \{1, \dots, p\}$. In order to study the consequences of this further, we offer some detailed examples in Section 5. More generally, the variance decay condition also arises for other forms of count data. For instance, in the case of a high-dimensional distribution with sparse Poisson marginals, the relation $\operatorname{var}(X_{i,j}) = \mathbb{E}[X_{i,j}]$ shows that weak sparsity in the mean vector can lead to variance decay.

- *Fourier coefficients of functional data.* Let $Y_1, \dots, Y_n$ be an i.i.d. sample of functional data, taking values in a separable Hilbert space $\mathcal{H}$. In addition, suppose that the covariance operator $\mathcal{C} = \operatorname{cov}(Y_1)$ is trace-class, which implies an eigenvalue decay condition of the form (2.7). Lastly, for each $i \in \{1, \dots, n\}$, let $X_i \in \mathbb{R}^p$ denote the first p generalized Fourier coefficients of Y_i with respect to some fixed orthonormal basis $\{\psi_j\}$ for $\mathcal{H}$. That is, $X_i = (\langle Y_i, \psi_1\rangle, \dots, \langle Y_i, \psi_p\rangle)$.

  Under the above conditions, it can be shown that no matter which basis $\{\psi_j\}$ is chosen, the vectors $X_1, \dots, X_n$ always satisfy the variance decay condition. (This follows from Proposition 2.1 below.) In Section 4, we explore some consequences of this condition as it relates to simultaneous confidence intervals for the Fourier coefficients of the mean function $\mathbb{E}[Y_1]$.

To conclude this section, we state a proposition that was used in the examples above. This basic result shows that decay among the eigenvalues $\lambda_1(\Sigma), \dots, \lambda_p(\Sigma)$ requires at least some decay among $\sigma_1, \dots, \sigma_p$.

PROPOSITION 2.1. *Fix two numbers $s \geq 1$ and $r \in (0, s)$. Then, there is a constant $c_{r,s} > 0$ depending only on r and s, such that for any symmetric matrix $A \in \mathbb{R}^{p\times p}$, we have*

$$\|\mathrm{diag}(A)\|_{w\ell_s} \leq c_{r,s}\|\lambda(A)\|_{w\ell_r}.$$

In particular, if $A = \Sigma$, and if there is a constant $c_0 > 0$ such that the inequality

$$\lambda_j(\Sigma) \leq c_0 j^{-1/r}$$

holds for all $1 \leq j \leq p$, then the inequality

$$\sigma_{(j)}^2 \leq c_0 c_{r,s} j^{-1/s}$$

holds for all $1 \leq j \leq p$.

The proof is given in Appendix A, and follows essentially from the Schur–Horn majorization theorem, as well as inequalities relating $\|\cdot\|_r$ and $\|\cdot\|_{w\ell_r}$.

3. Main results. In this section, we present our main results on Gaussian approximation and bootstrap approximation.

3.1. *Gaussian approximation.* Let $\tilde{S}_n \sim N(0, \Sigma)$ and define the Gaussian counterpart of the partially standardized statistic M (1.5) according to

$$\tilde{M} = \max_{1\leq j\leq p} \tilde{S}_{n,j}/\sigma_j^{\tau_n}. \tag{3.1}$$

Our first theorem shows that in the presence of variance decay, the distribution $\mathcal{L}(\tilde{M})$ can approximate $\mathcal{L}(M)$ at a nearly parametric rate in Kolmogorov distance. Recall that for any random variables U and V, this distance is given by $d_{\mathrm{K}}(\mathcal{L}(U), \mathcal{L}(V)) = \sup_{t\in\mathbb{R}} |\mathbb{P}(U \leq t) - \mathbb{P}(V \leq t)|$.

THEOREM 3.1 (Gaussian approximation). *Fix any number $\delta \in (0, 1/2)$, and suppose that Assumptions 2.1 and 2.2 hold. In addition, suppose that $\tau_n \in [0, 1)$ with $(1 - \tau_n)\sqrt{\log(n)} \gtrsim 1$. Then,*

$$d_{\mathrm{K}}\big(\mathcal{L}(M), \mathcal{L}(\tilde{M})\big) \lesssim n^{-\frac{1}{2}+\delta}. \tag{3.2}$$

Remarks. As a basic observation, note that the result handles the ordinary max statistic T as a special case with $\tau_n = 0$. In addition, it is worth emphasizing that the rate does not depend on the dimension p, or the variance decay parameter α, provided that it is positive. In this sense, the result shows that even a small amount of structure can have a substantial impact on Gaussian approximation (in relation to existing $n^{-1/6}$ rates that hold when $\alpha = 0$). Lastly, the reason for imposing the lower bound on $1 - \tau_n$ is that if τ_n quickly approaches 1 as $n \to \infty$, then the variances $\mathrm{var}(S_{n,j}/\sigma_j^{\tau_n})$ will also quickly approach 1, thus eliminating the beneficial effect of variance decay.

3.2. *Multiplier bootstrap approximation.* In order to define the multiplier bootstrap counterpart of $\tilde{M}$, first define the sample covariance matrix

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top, \tag{3.3}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Next, let $S_n^\star \sim N(0, \widehat{\Sigma}_n)$, and define the associated max statistic as

$$M^\star = \max_{1\leq j\leq p} S_{n,j}^\star/\widehat{\sigma}_j^{\tau_n}, \tag{3.4}$$

where $(\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_p^2) = \operatorname{diag}(\widehat{\Sigma}_n)$. In the exceptional case when $\widehat{\sigma}_j = 0$ for some j, the expression $S^\star_{n,j}/\widehat{\sigma}_j$ is understood to be 0. This convention is natural, because the event $S^\star_{n,j} = 0$ holds with probability 1, conditionally on $\widehat{\sigma}_j = 0$.

Remarks. The above description of $M^\star$ differs from some previous works insofar as we have suppressed the role of "multiplier variables," and have defined $S^\star_n$ as a sample from $N(0, \widehat{\Sigma}_n)$. From a mathematical standpoint, this is equivalent to the multiplier formulation (Chernozhukov, Chetverikov and Kato (2013)), where $S^\star_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i^\star (X_i - \bar{X})$ and $\xi_1^\star, \dots, \xi_n^\star$ are i.i.d. $N(0, 1)$ random variables, generated independently of X.

THEOREM 3.2 (Bootstrap approximation). *Fix any number $\delta \in (0, 1/2)$, and suppose the conditions of Theorem 3.1 hold. Then, there is a constant $c > 0$ not depending on n, such that the event*

$$d_{\mathrm{K}}\big(\mathcal{L}(\tilde{M}), \mathcal{L}\big(M^\star | X\big)\big) \le c n^{-\frac{1}{2}+\delta} \tag{3.5}$$

occurs with probability at least $1 - \frac{c}{n}$.

Remarks. At a high level, the proofs of Theorems 3.1 and 3.2 are based on the following observation: When the variance decay condition holds, there is a relatively small subset of $\{1, \dots, p\}$ that is likely to contain the maximizing index for M. In other words, if $\widehat{J} \in \{1, \dots, p\}$ denotes a random index satisfying $M = S_{n,\widehat{J}}/\sigma_{\widehat{J}}^{\tau_n}$, then the "effective range" of $\widehat{J}$ is fairly small. Although this situation is quite intuitive when the decay parameter α is large, what is more surprising is that the effect persists even for small values of α.

Once the maximizing index $\widehat{J}$ has been localized to a small set, it becomes possible to use tools that are specialized to the regime where $p \ll n$. For example, Bentkus' multivariate Berry–Esseen theorem (Bentkus (2003)) (cf. Lemma G.1) is helpful in this regard. Another technical aspect of the proofs worth mentioning is that they make essential use of the sharp constants in Rosenthal's inequality, as established in (Johnson, Schechtman and Zinn (1985)) (Lemma G.4).

4. Numerical illustration with functional data. Due to advances in technology and data collection, functional data have become ubiquitous in the past two decades, and statistical methods for their analysis have received growing interest. General references and surveys may be found in Ferraty and Vieu (2006), Horváth and Kokoszka (2012), Hsing and Eubank (2015), Ramsay and Silverman (2005), Wang, Chiou and Müller (2016).

The purpose of this section is to present an illustration of how the partially standardized statistic M and the bootstrap can be employed to do inference on functional data. More specifically, we consider a one-sample test for a mean function, which proceeds by constructing simultaneous confidence intervals (SCI) for its Fourier coefficients. With regard to our theoretical results, this is a natural problem for illustration, because the Fourier coefficients of functional data typically satisfy the variance decay condition (1.2), as explained in the third example of Section 2.2. Additional background and recent results on mean testing for functional data may be found in Benko, Härdle and Kneip (2009), Cao, Yang and Todem (2012), Choi and Reimherr (2018), Degras (2011), Horváth, Kokoszka and Reeder (2013), Zhang et al. (2019), Zheng, Yang and Härdle (2014), as well as the references therein.

4.1. *Tests for the mean function.* To set the stage, let $\mathcal{H}$ be a separable Hilbert space of functions, and let $Y \in \mathcal{H}$ be a random function with mean $\mathbb{E}[Y] = \mu$. Given a sample $Y_1, \dots, Y_n$ of i.i.d. realizations of Y, a basic goal is to test

$$H_0 : \mu = \mu^\circ \quad \text{versus} \quad H_1 : \mu \ne \mu^\circ, \tag{4.1}$$

where μ° is a fixed function in $\mathcal{H}$.

This testing problem can be naturally formulated in terms of SCI, as follows. Let $\{\psi_j\}$ denote any fixed orthonormal basis for $\mathcal{H}$. Also, let $\{u_j\}$ and $\{u_j^\circ\}$, respectively, denote the generalized Fourier coefficients of μ and μ° with respect to $\{\psi_j\}$, so that

$$\mu = \sum_{j=1}^{\infty} u_j \psi_j \quad \text{and} \quad \mu^\circ = \sum_{j=1}^{\infty} u_j^\circ \psi_j.$$

Then the null hypothesis is equivalent to $u_j = u_j^\circ$ for all $j \geq 1$. To test this condition, one can construct a confidence interval $\widehat{\mathcal{I}}_j$ for each u_j, and reject the null if $u_j^\circ \notin \widehat{\mathcal{I}}_j$ for at least one $j \geq 1$. In practice, due to the infinite dimensionality of $\mathcal{H}$, one will choose a sufficiently large integer p, and reject the null if $u_j^\circ \notin \widehat{\mathcal{I}}_j$ for at least one $j \in \{1, \ldots, p\}$.

Recently, a similar general strategy was pursued by Choi and Reimherr (2018), hereafter CR, who developed a test for the problem (4.1) based on a hyperrectangular confidence region for $(u_1, \ldots, u_p)$, which is equivalent to constructing SCI. In the CR approach, the basis is taken to be the eigenfunctions $\{\psi_{\mathcal{C},j}\}$ of the covariance operator $\mathcal{C} = \operatorname{cov}(Y)$, and p is chosen as the number of eigenfunctions $\psi_{\mathcal{C},1}, \ldots, \psi_{\mathcal{C},p}$ required to explain a certain fraction (say 99%) of variance in the data. However, since $\mathcal{C}$ is unknown, the eigenfunctions must be estimated from the available data.

When p is large, estimating the eigenfunctions $\psi_{\mathcal{C},1}, \ldots, \psi_{\mathcal{C},p}$ is a well-known challenge in functional data analysis. For instance, a large choice of p may be needed to explain 99% of the variance if the sample paths of $Y_1, \ldots, Y_n$ are not sufficiently smooth. Another example occurs when H_1 holds but μ and μ° are not well separated, which may require a large choice of p in order to distinguish $(u_1, \ldots, u_p)$ and $(u_1^\circ, \ldots, u_p^\circ)$. In light of these considerations, we will pursue an alternative approach to constructing SCI that does not require estimation of eigenfunctions.

4.2. *Applying the bootstrap.* Let $\{\psi_j\}$ be any pre-specified orthonormal basis for $\mathcal{H}$. For instance, when $\mathcal{H} = L^2[0, 1]$, a natural option is the standard Fourier basis. For a sample $Y_1, \ldots, Y_n \in \mathcal{H}$ as considered before, define random vectors $X_1, \ldots, X_n$ in $\mathbb{R}^p$ according to

$$X_i = \big(\langle Y_i, \psi_1 \rangle, \ldots, \langle Y_i, \psi_p \rangle\big),$$

and note that $\mathbb{E}[X_1] = (u_1, \ldots, u_p)$. For simplicity, we retain the previous notations associated with $X_1, \ldots, X_n$, so that $S_{n,j} = n^{-1/2} \sum_{i=1}^n (X_{i,j} - u_j)$, and likewise for other quantities. In addition, for any $\tau_n \in [0, 1]$, let

$$L = \min_{1 \leq j \leq p} S_{n,j} / \sigma_j^{\tau_n} \quad \text{and} \quad M = \max_{1 \leq j \leq p} S_{n,j} / \sigma_j^{\tau_n}.$$

For a given significance level $\varrho \in (0, 1)$, the ϱ-quantiles of L and M are denoted $q_L(\varrho)$ and $q_M(\varrho)$. Thus, the following event occurs with probability at least $1 - \varrho$:

$$\bigcap_{j=1}^{p} \left\{ \frac{q_L(\varrho/2)\sigma_j^{\tau_n}}{\sqrt{n}} \leq \bar{X}_j - u_j \leq \frac{q_M(1 - \varrho/2)\sigma_j^{\tau_n}}{\sqrt{n}} \right\}, \tag{4.2}$$

which leads to theoretical SCI for $(u_1, \ldots, u_p)$.

We now apply the bootstrap from Section 3.2 to estimate $q_L(\varrho/2)$ and $q_M(1 - \varrho/2)$. Specifically, we generate $B \geq 1$ independent samples of $M^\star$ as in (3.4), and then define $\widehat{q}_M(1 - \varrho/2)$ to be the empirical $(1 - \varrho/2)$-quantile of the B samples (and similarly for $\widehat{q}_L(\varrho/2)$), leading to the bootstrap SCI

$$\widehat{\mathcal{I}}_j = \left[\bar{X}_j - \frac{\widehat{q}_M(1 - \varrho/2)\widehat{\sigma}_j^{\tau_n}}{\sqrt{n}}, \bar{X}_j - \frac{\widehat{q}_L(\varrho/2)\widehat{\sigma}_j^{\tau_n}}{\sqrt{n}} \right] \tag{4.3}$$

for each $j \in \{1, \ldots, p\}$.

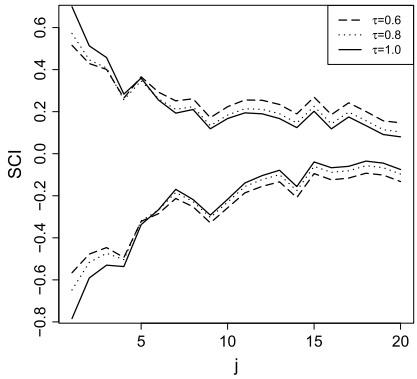

FIG. 1. *Illustration of the impact of τ_n on the shape of simultaneous confidence intervals* (*SCI*). *The curves represent upper and lower endpoints of the respective SCI, where the Fourier coefficients are indexed by j. Overall, the plot shows that the SCI change very gradually as a function of τ_n, and that there is a trade-off in the widths of the intervals. Namely, as τ_n decreases, the intervals for the leading coefficients* (*small j*) *become tighter, while the intervals for the subsequent coefficients* (*large j*) *become wider.*

It remains to select the value of τ_n, for which we adopt the following simple rule. For each choice of τ_n in a set of possible candidates, say $\mathcal{T} = \{0, 0.1, \ldots, 0.9, 1\}$, we construct the associated intervals $\widehat{\mathcal{I}}_1, \ldots, \widehat{\mathcal{I}}_p$ as in (4.3), and then select the value $\tau_n \in \mathcal{T}$ for which the average width $\frac{1}{p}\sum_{j=1}^{p} |\widehat{\mathcal{I}}_j|$ is the smallest, where $|[a, b]| = b - a$.

In Figure 1, we illustrate the influence of τ_n on the shape of the SCI. There are two main points to notice: (1) The intervals change very gradually as a function of τ_n, which shows that partial standardization is at most a mild adjustment of ordinary standardization. (2) The choice of τ_n involves a tradeoff, which controls the "allocation of power" among the p intervals. When τ_n is close to 1, the intervals are wider for the leading coefficients (small j), and narrower for the subsequent coefficients (large j). However, as τ_n decreases from 1, the widths of the intervals gradually become more uniform, and the intervals for the leading coefficients become narrower. Hence, if the vectors $(u_1, \ldots, u_p)$ and $(u_1^\circ, \ldots, u_p^\circ)$ differ in the leading coefficients, then choosing a smaller value of τ_n may lead to a gain in power. One last interesting point to mention is that in the simulations reported below, the selection rule of "minimizing the average width" typically selected values of τ_n around 0.8, and hence strictly less than 1.

4.3. *Simulation settings.* To study the numerical performance of the SCI described above, we generated i.i.d. samples from a Gaussian process on [0, 1], with population mean function

$$\mu_{\omega,\rho,\theta}(t) = (1+\rho)\cdot\big(\exp\big[-\{g_\omega(t)+2\}^2\big] + \exp\big[-\{g_\omega(t)-2\}^2\big]\big) + \theta$$

indexed by parameters (ω, ρ, θ), where $g_\omega(t) := 8h_\omega(t) - 4$, and $h_\omega(t)$ denotes the Beta distribution function with shape parameters $(2+\omega, 2)$. This family of functions was considered in Chen and Müller (2012). To interpret the parameters, note that ω determines the shape of the mean function (see Figure 2), whereas ρ and θ are scale and shift parameters. In terms of these parameters, the null hypothesis corresponds to $\mu = \mu^\circ := \mu_{0,0,0}$.

The population covariance function was taken to be the Matérn function

$$\mathcal{C}(s,t) = \frac{(\sqrt{2\nu}|t-s|)^\nu}{16\Gamma(\nu)2^{\nu-1}} K_\nu\big(\sqrt{2\nu}|t-s|\big),$$

which was previously considered in CR, with K_ν being a modified Bessel function of the second kind. We set $\nu = 0.1$, which results in relatively rough sample paths, as illustrated in

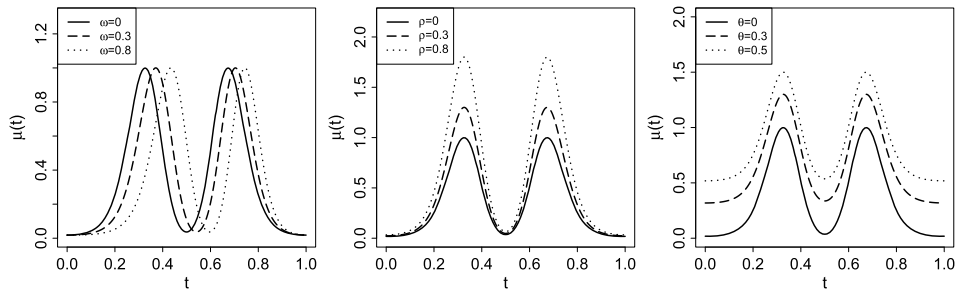

FIG. 2. *Left*: *Mean functions for varying shape parameters* ω *with* $\rho = \theta = 0$. *Middle*: *Mean functions for varying scale parameters* ρ *with* $\omega = \theta = 0$. *Right*: *Mean functions with different shift parameters* θ *with* $\omega = \rho = 0$.

the left panel of Figure 3. Also, the significant presence of variance decay is shown in the right panel.

When implementing the bootstrap in Section 4.2, we used the first $p = 100$ functions from the standard Fourier basis on [0,1]. (In principle, an even larger value p could have been selected, but we chose $p = 100$ to limit computation time.) For comparison purposes, we also implemented the "R_{Zs}" version of the method proposed in CR, using the accompanying R package `fregion` (Choi and Reimherr (2016)) under default settings, which typically utilized estimates of the first $p \approx 50$ eigenfunctions of $\mathcal{C}$.

Results on type I error. The nominal significance level was set to 5% in all simulations. To assess the actual type I error, we carried out 5000 simulations under the null hypothesis, for both $n = 50$ and $n = 200$. When $n = 50$, the type I error was 6.7% for the bootstrap method, and 1.6% for CR. When $n = 200$, the results were 5.7% for the bootstrap method, and 2.6% for CR. So, in these cases, the bootstrap respects the nominal significance level relatively well. In addition, our numerical results support the idea that partial standardization can be beneficial, because in the fully standardized case where $\tau_n = 1$, we observed less accurate type I error rates of 7.0% for $n = 50$, and 6.4% for $n = 200$.

Results on power. To consider power, we varied each of the parameters ω, ρ and θ, one at a time, while keeping the other two at their baseline value of zero. In each parameter setting, we carried out 1000 simulations with sample size $n = 50$. The results are summarized in Figure 4, showing that the bootstrap achieves relative gains in power—especially with respect to the shape (ω) and scale (ρ) parameters. In particular, it seems that using a large number of basis functions can help to catch small differences in these parameters (see also Figure 2).

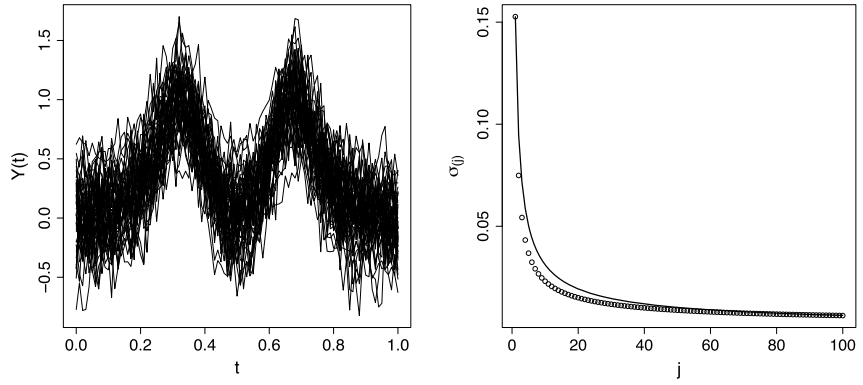

FIG. 3. *Left*: *A sample of the functional data* $Y_1, \ldots, Y_n$ *in the simulation study. Right*: *The ordered values* $\sigma_{(j)} = \sqrt{\text{var}(X_{1,j})}$ *are represented by dots, which are approximated by the decay profile* $0.15\, j^{-0.69}$ (*solid line*).

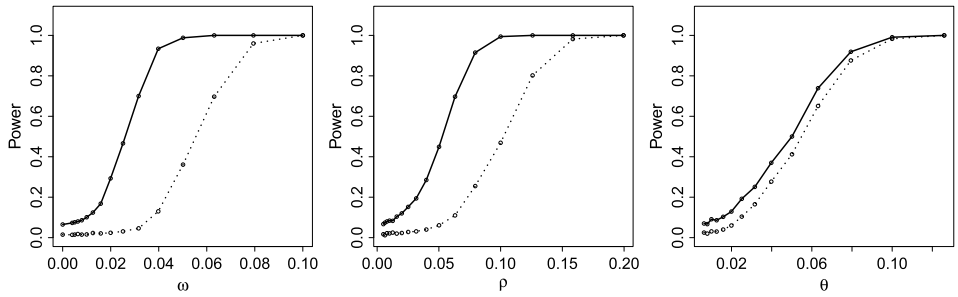

FIG. 4. *Empirical power for the partially standardized bootstrap method* (*solid*) *and the CR method* (*dotted*) *Left*: *Empirical power for varying shape parameters* ω *while* $\rho = \theta = 0$. *Middle*: *Empirical power for varying scale parameters* ρ *while* $\omega = \theta = 0$. *Right*: *Empirical power for varying shift parameters* θ *while* $\omega = \rho = 0$.

5. Examples with multinomial data. When multinomial models are used in practice, it is not uncommon for the number of cells p to be quite large. Indeed, the challenges of this situation have been a topic of sustained interest, and many inferential questions remain unresolved (e.g., Balakrishnan and Wasserman (2019), Chafaï and Concordet (2009), Cressie and Read (1984), Fienberg and Holland (1973), Hoeffding (1965), Holst (1972), Paninski (2008), Zelterman (1987)). A recent survey is (Balakrishnan and Wasserman (2018)). As one illustration of how our approach can be applied to such models, this section will look at the task of constructing SCI for the cell proportions. Although this type of problem has been studied from a variety of perspectives over the years (e.g., Chafaï and Concordet (2009), Fitzpatrick and Scott (1987), Goodman (1965), Quesenberry and Hurst (1964), Sison and Glaz (1995), Wang (2008)), relatively few theoretical results directly address the high-dimensional setting—and in this respect, our example offers some progress. Lastly, it is notable that multinomial data are of a markedly different character than the functional data considered in Section 4, which demonstrates how our approach has a broad scope of potential applications.

5.1. *Theoretical example.* Recall from Section 2.2 that we regard the observations in the multinomial model as lying in the set of standard basis vectors $\{e_1, \dots, e_p\} \subset \mathbb{R}^p$. In this context, we also write $\widehat{\pi}_j = \bar{X}_j$ to indicate that the jth coordinate of the sample mean is an estimate of the jth cell proportion π_j. In addition, it is important to clarify that a variance decay condition of the form (1.2) is automatically satisfied in this model (as explained in Section 2.2), and so it is not necessary to include this as a separate assumption. Below, we retain the definition of k_n in (2.2).

ASSUMPTION 5.1 (Multinomial model).

(i) The observations $X_1, \dots, X_n \in \mathbb{R}^p$ are i.i.d., with $\mathbb{P}(X_1 = e_j) = \pi_j$ for each $j \in \{1, \dots, p\}$, where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$ is a probability vector that may vary with n.

(ii) There are constants $\alpha > 0$ and $\epsilon_0 \in (0, 1)$, with neither depending on n, such that

$$\sigma_{(j)} \geq \epsilon_0 j^{-\alpha} \quad \text{for all } j \in \{1, \dots, (k_n + 1) \wedge p\}. \tag{5.1}$$

Remarks. A concrete set of examples satisfying the conditions of Assumption 5.1 is given by probability vectors of the form $\pi_{(j)} \propto j^{-\eta}$, with $\eta > 1$. Furthermore, the condition $\eta > 1$ is mild, since the inequality $\pi_{(j)} \leq j^{-1}$ is satisfied by every probability vector.

Applying the bootstrap. In the high-dimensional setting, the multinomial model differs in an essential way from the model in Section 2, because there will often be many empty cells

(indices) $j \in \{1, \ldots, p\}$ for which $\widehat{\sigma}_j = 0$. For the indices where this occurs, the usual confidence intervals of the form (4.3) have zero width, and thus cannot be used. More generally, if the number of observations in cell j is small, then it is inherently difficult to construct a good confidence interval around π_j. Consequently, we will restrict our previous SCI (4.3) by focusing on a set of cells that contain a sufficient number of observations. For theoretical purposes, such a set may be defined as

$$\widehat{\mathcal{J}}_n = \left\{ j \in \{1, \ldots, p\} \,\middle|\, \widehat{\pi}_j \geq \sqrt{\frac{\log(n)}{n}} \right\}. \tag{5.2}$$

Accordingly, the max statistic and its bootstrapped version are defined by taking maxima over the indices in $\widehat{\mathcal{J}}_n$, and we denote them as

$$\mathcal{M} = \max_{j \in \widehat{\mathcal{J}}_n} S_{n,j}/\sigma_j^{\tau_n}$$

and

$$\mathcal{M}^\star = \max_{j \in \widehat{\mathcal{J}}_n} S_{n,j}^\star/\widehat{\sigma}_j^{\tau_n},$$

where we arbitrarily take $\mathcal{M}$ and $\mathcal{M}^\star$ to be zero in the exceptional case when $\widehat{\mathcal{J}}$ is empty.

Although the presence of the random index set $\widehat{\mathcal{J}}_n$ complicates the distributions of $\mathcal{M}$ and $\mathcal{M}^\star$, it is a virtue of the bootstrap that this source of randomness is automatically accounted for in the resulting inference. In addition, the following result shows that the bootstrap continues to achieve a near-parametric rate of approximation.

THEOREM 5.1. *Fix any $\delta \in (0, 1/2)$, and suppose that Assumption 5.1 holds. In addition, suppose that $\tau_n \in [0, 1)$ with $(1 - \tau_n)\sqrt{\log(n)} \gtrsim 1$. Then there is a constant $c > 0$ not depending on n such that the event*

$$d_{\mathrm{K}}\big(\mathcal{L}(\mathcal{M}), \mathcal{L}\big(\mathcal{M}^\star | X\big)\big) \leq cn^{-1/2+\delta}, \tag{5.3}$$

occurs with probability at least $1 - \frac{c}{n}$.

Remarks. The proof of this result shares much of the same structure as the proofs of Theorems 3.1 and 3.2, but there are a few differences. First, the use of the random index set $\widehat{\mathcal{J}}_n$ in the definition of $\mathcal{M}$ and $\mathcal{M}^\star$ entails some extra technical considerations, which are handled with the help of Kiefer's inequality (Lemma G.5). Second, we develop a lower bound for $\lambda_{\min}(\Sigma(k_n))$, where $\Sigma(k_n)$ is the covariance matrix of the variables indexed by $\mathcal{J}(k_n)$ (see Lemma F.3). This bound may be of independent interest for problems involving multinomial distributions, and does not seem to be well known; see also (Bénasséni (2012)) for other related eigenvalue bounds.

5.2. *Numerical example.* We illustrate the bootstrap procedure in the case of the model $\pi_j \propto j^{-1}$, which was considered in a recent numerical study of Balakrishnan and Wasserman (2018). Taking $p = 1000$ and $n \in \{500, 1000\}$, we applied the bootstrap method to construct 95% SCI for the proportions π_j corresponding to the cells with at least 5 observations. The cutoff value of 5 is based on a guideline that is commonly recommended in textbooks, for example, Agresti (2002), page 19, Rice (2007), page 519. Lastly, the parameter τ_n was chosen in the same way as described in Section 4.2.

Based on 5000 Monte Carlo runs, the observed coverage probability was found to be 93.7% for $n = 500$, and 94.4% for $n = 1000$, demonstrating satisfactory performance. Regarding the parameter τ_n, the selection rule typically produced values close to 0.8, for both

$n = 500$ and $n = 1000$. As a point of comparison, it is also interesting to mention the coverage probabilities that occurred when τ_n was set to 1 (which eliminates all variance decay). In this case, the coverage probabilities became less accurate, with values of 92.7% for $n = 500$, and 93.1% for $n = 1000$. Hence, this shows that taking advantage of variance decay can enhance coverage probability.

6. Conclusions. The main conclusion to draw from our work is that a modest amount of variance decay in a high-dimensional model can substantially improve rates of bootstrap approximation for max statistics, which helps to reconcile some of the empirical and theoretical results in the literature. In particular, there are three aspects of this type of model structure that are worth emphasizing. First, the variance decay condition (1.2) is very weak, in the sense that the parameter $\alpha > 0$ is allowed to be arbitrarily small. Second, the condition is approximately checkable in practice, since the parameters $\sigma_1, \ldots, \sigma_p$ can be accurately estimated when $n \ll p$. Third, this type of structure arises naturally in a variety of contexts.

Beyond our main theoretical focus on rates of bootstrap approximation, we have also shown that the technique of partial standardization leads to favorable numerical results. Specifically, this was illustrated with examples involving both functional and multinomial data, where variance decay is an inherent property that can be leveraged. Finally, we note that these applications are by no means exhaustive, and the adaptation of the proposed approach to other types of data may provide further opportunities for future work.

Acknowledgments. The first author was supported in part by NSF Grant DMS-1613218.
The second author was supported in part by NIH Grant 5UG3OD023313-03.
The third author was supported in part by NSF Grant DMS-1712864 and NIH Grant 5UG3OD023313-03.

SUPPLEMENTARY MATERIAL

Supplement to "Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data" (DOI: 10.1214/19-AOS1844SUPP; .pdf). The supplement contains proofs of all theoretical results.

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, New York. MR1914507 https://doi.org/10.1002/0471249688

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010a). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.* **38** 51–82. MR2589316 https://doi.org/10.1214/08-AOS667

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010b). Some nonasymptotic results on resampling in high dimension. II. Multiple tests. *Ann. Statist.* **38** 83–99. MR2589317 https://doi.org/10.1214/08-AOS668

BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.* **12** 727–749. MR3834283 https://doi.org/10.1214/18-AOAS1155SF

BALAKRISHNAN, S. and WASSERMAN, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.* **47** 1893–1927. MR3953439 https://doi.org/10.1214/18-AOS1729

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D., HANSEN, C. and KATO, K. (2018). High-dimensional econometrics and regularized GMM. arXiv:1806.01888.

BÉNASSÉNI, J. (2012). A new derivation of eigenvalue inequalities for the multinomial distribution. *J. Math. Anal. Appl.* **393** 697–698. MR2921708 https://doi.org/10.1016/j.jmaa.2012.03.029

BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343 https://doi.org/10.1214/07-AOS516

BENTKUS, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *J. Statist. Plann. Inference* **113** 385–402. MR1965117 https://doi.org/10.1016/S0378-3758(02)00094-0

BUNEA, F. and XIAO, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli* **21** 1200–1230. MR3338661 https://doi.org/10.3150/14-BEJ602

CAO, G., YANG, L. and TODEM, D. (2012). Simultaneous inference for the mean function based on dense functional data. *J. Nonparametr. Stat.* **24** 359–377. MR2921141 https://doi.org/10.1080/10485252.2011.638071

CHAFAÏ, D. and CONCORDET, D. (2009). Confidence regions for the multinomial parameter with small sample size. *J. Amer. Statist. Assoc.* **104** 1071–1079. MR2562005 https://doi.org/10.1198/jasa.2009.tm08152

CHANG, J., YAO, Q. and ZHOU, W. (2017). Testing for high-dimensional white noise using maximum cross-correlations. *Biometrika* **104** 111–127. MR3626482 https://doi.org/10.1093/biomet/asw066

CHEN, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Ann. Statist.* **46** 642–678. MR3782380 https://doi.org/10.1214/17-AOS1563

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2015). Asymptotic theory for density ridges. *Ann. Statist.* **43** 1896–1928. MR3375871 https://doi.org/10.1214/15-AOS1329

CHEN, D. and MÜLLER, H.-G. (2012). Nonlinear manifold representations for functional data. *Ann. Statist.* **40** 1–29. MR3013177 https://doi.org/10.1214/11-AOS936

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468 https://doi.org/10.1214/14-AOS1235

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963 https://doi.org/10.1214/16-AOP1113

CHOI, H. and REIMHERR, M. (2016). R package '`fregion`'. https://github.com/hpchoi/fregion.

CHOI, H. and REIMHERR, M. (2018). A geometric approach to confidence regions and bands for functional parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 239–260. MR3744720 https://doi.org/10.1111/rssb.12239

CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464. MR0790631

DEGRAS, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statist. Sinica* **21** 1735–1765. MR2895997 https://doi.org/10.5705/ss.2009.207

DENG, H. and ZHANG, C. H. (2017). Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. arXiv:1705.09528.

DEZEURE, R., BÜHLMANN, P. and ZHANG, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719. MR3713586 https://doi.org/10.1007/s11749-017-0554-2

FAN, J., SHAO, Q.-M. and ZHOU, W.-X. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *Ann. Statist.* **46** 989–1017. MR3797994 https://doi.org/10.1214/17-AOS1575

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. *Springer Series in Statistics*. Springer, New York. MR2229687

FIENBERG, S. E. and HOLLAND, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68** 683–691. MR0359153

FITZPATRICK, S. and SCOTT, A. (1987). Quick simultaneous confidence intervals for multinomial proportions. *J. Amer. Statist. Assoc.* **82** 875–878. MR0909995

GOODMAN, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7** 247–254.

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **36** 369–408. MR0173322 https://doi.org/10.1214/aoms/1177700150

HOLST, L. (1972). Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika* **59** 137–145. MR0314193 https://doi.org/10.1093/biomet/59.1.137

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. MR2920735 https://doi.org/10.1007/978-1-4614-3655-3

HORVÁTH, L., KOKOSZKA, P. and REEDER, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 103–122. MR3008273 https://doi.org/10.1111/j.1467-9868.2012.01032.x

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

JOHNSON, W. B., SCHECHTMAN, G. and ZINN, J. (1985). Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *Ann. Probab.* **13** 234–253. MR0770640

JUNG, S., LEE, M. H. and AHN, J. (2018). On the number of principal components in high dimensions. *Biometrika* **105** 389–402. MR3804409 https://doi.org/10.1093/biomet/asy010

KOLTCHINSKII, V., LÖFFLER, M. and NICKL, R. (2020). Efficient estimation of linear functionals of principal components. *Ann. Statist.* **48** 464–490. MR4065170 https://doi.org/10.1214/19-AOS1816

KOLTCHINSKII, V. and LOUNICI, K. (2017a). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

KOLTCHINSKII, V. and LOUNICI, K. (2017b). Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.* **45** 121–157. MR3611488 https://doi.org/10.1214/16-AOS1437

LOPES, M. E., LIN, Z. and MÜLLER, H.-G. (2020). Supplement to "Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data." https://doi.org/10.1214/19-AOS1844SUPP.

LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437 https://doi.org/10.3150/12-BEJ487

NAUMOV, A., SPOKOINY, V. and ULYANOV, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probab. Theory Related Fields* **174** 1091–1132. MR3980312 https://doi.org/10.1007/s00440-018-0877-2

PANINSKI, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory* **54** 4750–4755. MR2591136 https://doi.org/10.1109/TIT.2008.928987

QUESENBERRY, C. P. and HURST, D. C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* **6** 191–195. MR0184329 https://doi.org/10.2307/1266151

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

REISS, M. and WAHL, M. (2020). Nonasymptotic upper bounds for the reconstruction error of PCA. *Ann. Statist.* **48** 1098–1123.

RICE, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury Press, Pacific Grove CA.

SISON, C. P. and GLAZ, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *J. Amer. Statist. Assoc.* **90** 366–369. MR1325142

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

WANG, H. (2008). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *J. Multivariate Anal.* **99** 896–911. MR2405097 https://doi.org/10.1016/j.jmva.2007.05.003

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

WASSERMAN, L., KOLAR, M. and RINALDO, A. (2014). Berry–Esseen bounds for estimating undirected graphs. *Electron. J. Stat.* **8** 1188–1224. MR3263117 https://doi.org/10.1214/14-EJS928

ZELTERMAN, D. (1987). Goodness-of-fit tests for large sparse multinomial distributions. *J. Amer. Statist. Assoc.* **82** 624–629. MR0898368

ZHANG, X. and CHENG, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768. MR3671768 https://doi.org/10.1080/01621459.2016.1166114

ZHANG, J.-T., CHENG, M.-Y., WU, H.-T. and ZHOU, B. (2019). A new test for functional one-way ANOVA with applications to ischemic heart screening. *Comput. Statist. Data Anal.* **132** 3–17. MR3913131 https://doi.org/10.1016/j.csda.2018.05.004

ZHENG, S., YANG, L. and HÄRDLE, W. K. (2014). A smooth simultaneous confidence corridor for the mean of sparse functional data. *J. Amer. Statist. Assoc.* **109** 661–673. MR3223741 https://doi.org/10.1080/01621459.2013.866899