Statistica Sinica Preprint No: SS-2025-0036	
Title	Robust Max Statistics for High-dimensional Inference
Manuscript ID	SS-2025-0036
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0036
Complete List of Authors	Mingshuo Liu and Miles Lopes
Corresponding Authors	Miles Lopes
E-mails	melopes@ucdavis.edu
Notice: Accepted author version.	

Robust Max Statistics for High-Dimensional Inference

Mingshuo Liu and Miles E. Lopes

University of California, Davis

Abstract:

Although much progress has been made in the theory and application of bootstrap approximations for max statistics in high dimensions, the literature has largely been restricted to cases involving light-tailed data. To address this issue, we propose an approach to inference based on *robust max statistics*, and we show that their distributions can be accurately approximated via bootstrapping when the data are both high-dimensional and heavy-tailed. In particular, the data are assumed to satisfy an extended version of the well-established L^4-L^2 moment equivalence condition, as well as a weak variance decay condition. In this setting, we show that *near-parametric* rates of bootstrap approximation can be achieved in the Kolmogorov metric, *independently of the data dimension.* Moreover, this theoretical result is complemented by encouraging empirical results involving both Euclidean and functional data.

Key words and phrases: high-dimensional statistics; robustness; bootstrap; simultaneous inference; median-of-means

1. Introduction

Over the past decade, distributional approximation results for max statistics have become a prominent topic in high-dimensional inference. A prototypical example of such a statistic has the form $\max_{1\le j\le p}\sqrt{n}(\bar{X}_j-\mu_j)$, where $\bar{X}\in\mathbb{R}^p$ is the sample mean vector of n obser-

vations and $\mu = \mathbf{E}(\bar{X})$, but numerous variants arise in diverse contexts. Indeed, one of the main drivers of research on this topic is that many high-dimensional inference tasks can be unified within the problem of approximating the distribution of $\max_{1\leq j\leq p}\sqrt{n}(\bar{X}_j - \mu_j)$, or some adaptation of it. For instance, such approximations can be directly applied to construct simultaneous tests and confidence intervals for coordinate-wise means $\mu_1, \ldots, \mu_p$. More broadly, other applications include detection of treatment effects (Sun et al., 2022), error estimation for sample covariance matrices (Lopes et al., 2023), post-selection inference (Kuchibhotla et al., 2020), change-point detection (Yu and Chen, 2021), confidence bands in non-parametric regression (Singh and Vijaykumar, 2023), tests for shape restrictions (Chetverikov et al., 2018), and more. Meanwhile, another major reason why max statistics have attracted growing interest is that bootstrap methods can accurately approximate their distributions when p is much larger than the sample size n, which has been demonstrated by a cascade of theoretical advances (Chernozhukov et al., 2013, 2017; Deng and Zhang, 2020; Kuchibhotla and Rinaldo, 2020; Lopes et al., 2020; Kuchibhotla et al., 2021; Lopes, 2022; Chernozhukov et al., 2023; Fang et al., 2023; Koike, 2024).

Despite the substantial innovations that have been made in bootstrap approximations for max statistics, there is an Achilles heel that continues to hinder much of the research in this area. Namely, there is a widespread reliance on the assumption that the covariates have light tails, e.g. sub-Gaussian or sub-exponential. Moreover, there are empirical and theoretical results suggesting that light tails *are necessary* for bootstrap methods to successfully approximate the distributions of conventional max statistics in high dimensions (Zhang and Wu, 2017; Giessing and Fan, 2020; Kock and Preinerstorfer,

2024). For instance, the simulations in (Giessing and Fan, 2020) show that the Gaussian multiplier bootstrap performs poorly for $\max_{1\le j\le p}\sqrt{n}|\bar{X}_j-\mu_j|$ when the covariates have heavy tails and $p \gg n$. From a theoretical standpoint, it has also been proven that there is a moment-dependent phase transition governing the success of Gaussian approximations for $\max_{1\le j\le p}\sqrt{n}(\bar{X}_j-\mu_j)$ (Zhang and Wu, 2017; Kock and Preinerstorfer, 2024). That is, if $W \in \mathbb{R}^p$ is a centered Gaussian vector having the same covariance matrix as $\sqrt{n}(\bar{X}-\mu)$, then the Kolmogorov distance between $\max_{1\le j\le p} W_j$ and $\max_{1\le j\le p}\sqrt{n}(\bar{X}_j-\mu_j)$ may or may not vanish in the limit that n and p jointly diverge, depending on whether the covariates have enough moments. This breakdown of Gaussian approximations suggests that similar behavior should occur for bootstrap approximations—especially in the case of the Gaussian multiplier bootstrap, which seeks to mimic the behavior of $\sqrt{n}(\bar{X}-\mu)$ by generating random vectors from a centered Gaussian distribution whose covariance matrix is an estimate for that of $\sqrt{n}(\bar{X}-\mu)$.

1.1 Existing approaches for addressing heavy tails

Due to the issues just mentioned, there are strong motivations to extend bootstrap methods involving max statistics so that they can be applied reliably to high-dimensional data with heavy tails. However, the research in this direction is still at an early stage, and there are only a few previous works that have given it attention. The first of these works briefly outlined an approach that combines truncation with permutation-based sampling (Lou and Wu, 2017), but it was ultimately not pursued as a practical method for inference. More recently, the paper (Fan et al., 2023) proposed a weighted bootstrap for a max statistic of the form $\max_{1\le j\le p}\sqrt{n}|\hat{\theta}_j-\theta_j|$, where θ_j denotes the so-called "pseudomedian"

of the jth covariate, and $\hat{\theta}_j$ is the classical Hodges-Lehmann estimator for θ_j (Hodges and Lehmann, 1963).

While Fan et al. (2023) achieved major progress by delivering robust simultaneous inference for $\theta_1, \ldots, \theta_p$, it still has some essential limitations. One is that the pseudomedians can be unsatisfactory substitutes for the means $\mu_1, \ldots, \mu_p$, particularly in cases of asymmetric distributions, for which θ_j and μ_j may be quite different. A related issue is that an approach based on pseudomedians does not extend naturally to suprema of zero-mean empirical processes, which appear frequently in applications of bootstrap approximations for max statistics (Chernozhukov et al., 2014; Chen et al., 2015; Han et al., 2018; Chen and Kato, 2020; Dette et al., 2020; Lopes et al., 2023; Giessing, 2023). Another issue is that the method in Fan et al. (2023) produces simultaneous confidence intervals for $\theta_1, \ldots, \theta_p$ that are only theoretically justified when they all have the same width, which is impractical if the covariates fluctuate over different scales. Lastly, the available theoretical analysis for $\max_{1 \leq j \leq p} \sqrt{n}|\hat{\theta}_j - \theta_j|$ establishes a near $n^{-1/4}$ rate of bootstrap approximation in the Kolmogorov metric, which does not align with other recent results for max statistics that establish near $n^{-1/2}$ rates in the setting of light-tailed data (Lopes et al., 2020; Lopes, 2022; Chernozhukov et al., 2023; Fang et al., 2023; Koike, 2024).

After the preprint version of this work was posted (Liu and Lopes, 2024), two papers by Resende (2024) and Kock and Preinerstorfer (2025) appeared, which proposed bootstrapping max statistics based on trimmed and truncated (Winsorized) sample means. The theoretical results in these works share a key limitation with those of Fan et al. (2023), insofar as the established rates of bootstrap approximation in the Kolmogorov

metric are no faster than $n^{-1/4}$, regardless of how many moments the data are assumed to possess. In addition to trimming and truncation, another possible way to enhance the robustness of sample means is self-normalization (de la Peña et al., 2009). In essence, adapting this technique to our context amounts to replacing the coordinate-wise sample means in the statistic $\max_{1\le j\le p}\sqrt{n}(\bar{X}_j-\mu_j)$ with coordinate-wise t-statistics, or variants thereof. Although the effectiveness of bootstrapping certain self-normalized statistics has been proven theoretically in various other high-dimensional inference tasks (e.g. Fan et al., 2007; Delaigle et al., 2011; Liu and Shao, 2014; Chang et al., 2016), rates of bootstrap approximation for self-normalized max statistics in the high-dimensional setting are not available.

1.2 Robust max statistics

In the current paper, we propose to bootstrap a robust max statistic that enables simultaneous inference on the means $\mu_1,\dots,\mu_p$ and overcomes many of the difficulties described above. Our approach is designed in terms of three ingredients: truncation, partial standardization, and the median-of-means (MOM) technique (Nemirovsky and Yudin, 1983; Lugosi and Mendelson, 2019). To briefly lay out the main ideas, let $X_1,\dots,X_n\in\mathbb{R}^p$ be i.i.d. observations with $\sigma_j^2=\mathrm{var}(X_{1j})$ and $\mu_j=\mathbf{E}(X_{1j})$ as before. Also, let $\hat{\sigma}_1^2,\dots,\hat{\sigma}_p^2$ denote variance estimates that will be constructed from a small hold-out set via MOM, and define the truncation function $\varphi_t(x)=\mathrm{sgn}(x)(|x|\wedge t)$ for any $x\in\mathbb{R}$ and $t\ge 0$, where $a\wedge b=\min\{a,b\}$. In this notation, the proposed robust max statistic is defined by

$$\mathcal{M}_n=\max_{1\le j\le p}\sum_{i=1}^{n}\frac{\varphi_{\hat{t}_j}(X_{ij}-\mu_j)}{\hat{\sigma}_j^{\tau}n^{1/2}}, \tag{1.1}$$

where $\hat{t}_j = \sqrt{n}\hat{\sigma}_j$ for $j = 1, \ldots, p$, and $\tau \in [0, 1]$ is a fixed partial standardization parameter.

Importantly, there is a direct link between distributional approximations for $\mathcal{M}_n$ and inference on the means $\mu_1, \ldots, \mu_p$. This is due to the monotonicity of the functions $\varphi_{\hat{t}_j}(\cdot)$, which makes it straightforward to construct simultaneous confidence intervals for the means using quantile estimates for $\mathcal{M}_n$ and its corresponding min statistic, as discussed in Section 2. The robust variance estimates $\hat{\sigma}_j^2$ also play an essential role, because they ensure that the confidence intervals induced by $\mathcal{M}_n$ are automatically adapted to the scale of the covariates, which is an issue that has often been neglected in the literature on max statistics.

For the purpose of bootstrapping $\mathcal{M}_n$, let $\tilde{X}_j$ denote a hold-out MOM estimate of μ_j to be defined later, and let $\bar{\varphi}_j = \frac{1}{n}\sum_{i=1}^n \varphi_{\hat{t}_j}(X_{ij} - \tilde{X}_j)$. In addition, let $\xi_1, \ldots, \xi_n \sim N(0, 1)$ be i.i.d. Gaussian multipliers generated independently of the data. Putting these pieces together, we define a bootstrap sample of $\mathcal{M}_n$ as

$$\mathcal{M}_n^* = \max_{1 \le j \le p} \sum_{i=1}^n \frac{\xi_i\big(\varphi_{\hat{t}_j}(X_{ij} - \tilde{X}_j) - \bar{\varphi}_j\big)}{\hat{\sigma}_j^\tau n^{1/2}}. \tag{1.2}$$

With regard to theoretical analysis, we focus on a setting where the tails of the data are quantified by a variant of the L^4-L^2 moment equivalence condition, which has gained increasing currency in the high-dimensional robustness literature (Lugosi and Mendelson, 2019; Ke et al., 2019; Mendelson and Zhivotovskiy, 2020; Roy et al., 2021; Abdalla and Zhivotovskiy, 2022). Specifically, we assume there is some $\delta > 0$ such that the bound $\|\langle v, X_1 - \mu\rangle\|_{L^{4+\delta}} \lesssim \|\langle v, X_1 - \mu\rangle\|_{L^2}$ holds for all $v \in \mathbb{R}^p$, and in Proposition 1, we show that this condition is satisfied by heavy-tailed instances of well-known models. The

other primary structural assumption in our analysis is that the covariates have a weak variance decay property of the form $\sigma^2_{(j)} \asymp j^{-2\beta}$ for some fixed $\beta > 0$, where $\sigma^2_{(1)} \geq \cdots \geq \sigma^2_{(p)}$ are the sorted coordinate-wise variances. Notably, the decay is referred to as weak because β is allowed to be *arbitrarily small*, and hence it is mild structural assumption. Furthermore, it is known that this type of structure arises naturally in a variety of high-dimensional contexts that are related to principal components analysis and functional data analysis, among others (Lopes et al., 2020). Under the complete set of conditions given in Assumption 1 later on, our main result shows that with high probability, the Kolmogorov distance $\sup_{s\in\mathbb{R}} |\mathbf{P}(\mathcal{M}_n \leq s) - \mathbf{P}(\mathcal{M}^*_n \leq s|X)|$ is nearly of order $n^{-1/2}$, where $\mathbf{P}(\cdot|X)$ denotes probability that is conditional on all of the observations.

From a practical standpoint, the proposed method has several strengths. First, the method does not require fine tuning, which is demonstrated by the fact that we use the simple choices of $\hat{t}_j = \sqrt{n}\hat{\sigma}_j$ and $\tau = 0.9$ throughout all of the experiments presented in Section 4. Second, we show that the method reliably produces well-calibrated tests and confidence intervals across many conditions—including heavy-tailed data generated from separable and elliptical distributions, as well as heavy-tailed functional data with rough sample paths. Third, the empirical performance of the proposed method is favorable in comparison to the pseudomedian approach in (Fan et al., 2023), as well as the trimming and the truncation approaches in (Kock and Preinerstorfer, 2025). (The trimming approach in (Resende, 2024) was developed at a theoretical level, and certain implementation details were not covered in enough detail to allow for empirical comparisons.) In addition, we include empirical results for a self-normalization approach that we adapted from the

paper (Liu and Shao, 2014), which focused on a set of inference problems different from those considered here. The self-normalization approach provides somewhat tighter confidence intervals than the proposed method while maintaining similar level control, but it is not supported by comparable theoretical guarantees in our context. A brief summary of the details involved in using these various alternative methods is provided in Section 4.

Notation. If A is a real matrix, its Frobenius norm is $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$, and its operator norm $\|A\|_{\mathrm{op}}$ is the same as its largest singular value. If x and y are Euclidean vectors of the same dimension, then $\langle x, y\rangle$ denotes the Euclidean inner product, and $\|x\|_2 = \sqrt{\langle x, x\rangle}$. If ξ is a scalar random variable and $1 \le q < \infty$, we write $\|\xi\|_{L^q} = (\mathbf{E}|\xi|^q)^{1/q}$, and in the case when $q = \infty$, we use $\|\xi\|_{L^\infty}$ to refer to the essential supremum. If f is a scalar-valued function on $\mathbb{R}$, the notation $\|f\|_{L^\infty}$ is understood analogously with respect to Lebesgue measure. If $\{a_n\}$ and $\{b_n\}$ are sequences of non-negative real numbers, then the relations $a_n \lesssim b_n$ and $a_n = \mathcal{O}(b_n)$ are equivalent, and mean that there is a constant $c > 0$ not depending on n, such that $a_n \le c b_n$ holds for all large n. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$ both hold, then we write $a_n \asymp b_n$. Lastly, let $a_n \vee b_n = \max\{a_n, b_n\}$.

2. Method

Here, we provide the details for constructing the bootstrap sample $\mathcal{M}_n^*$, as well as simultaneous confidence intervals $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$ for the coordinate-wise means $\mu_1, \ldots, \mu_p$. Further applications of these intervals to various testing problems will be covered later in Section 4.

In addition to the observations $X_1, \ldots, X_n$ discussed above, let $X_{n+1}, \ldots, X_{n+m_n}$ denote an independent set of i.i.d. hold-out observations generated from the same distri-

bution. For simplicity, the number of hold-out observations m_n is assumed to be even, and in all of our numerical experiments, we will take m_n to be about 10% of n. The hold-out observations are used to construct robust estimators $\tilde{X}_1, \ldots, \tilde{X}_p$ and $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2$ for the coordinate-wise means and variances, which are the only ingredients for generating $\mathcal{M}_n^*$ that were not addressed previously in Section 1. Taking an MOM approach, we partition the hold-out indices $\{n+1, \ldots, n+m_n\}$ into b_n blocks $\mathcal{B}_1, \ldots, \mathcal{B}_{b_n}$, with each block containing an even number of ℓ_n indices such that $m_n = \ell_n b_n$. More specifically, let $\mathcal{B}_1 = \{n+1, \ldots, n+\ell_n\}$, $\mathcal{B}_2 = \{n+\ell_n+1, \ldots, n+2\ell_n\}$, and so on. For the lth block, let

$$\bar{X}_j(l) = \frac{1}{\ell_n} \sum_{i \in \mathcal{B}_l} X_{ij} \tag{2.3}$$

denote the block-wise sample mean of the jth coordinate, and define the MOM estimator of μ_j as

$$\tilde{X}_j = \text{median}(\bar{X}_j(1), \ldots, \bar{X}_j(b_n)). \tag{2.4}$$

Likewise, we construct an MOM estimate for each σ_j^2 along similar lines. The lth block-wise estimate for σ_j^2 is obtained by averaging the squared differences of $\ell_n/2$ pairs of observations

$$\bar{\sigma}_j^2(l) = \frac{1}{\ell_n/2} \sum_{\substack{i, i' \in \mathcal{B}_l \\ i' - i = \ell_n/2}} \frac{1}{2}(X_{ij} - X_{i'j})^2, \tag{2.5}$$

and then $\hat{\sigma}_j^2$ is defined to be the median of the block-wise estimates

$$\hat{\sigma}_j^2 = \text{median}(\bar{\sigma}_j^2(1), \ldots, \bar{\sigma}_j^2(b_n)). \tag{2.6}$$

Next, we turn to the construction of simultaneous confidence intervals $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$ for $\mu_1, \ldots, \mu_p$. Let $1-\alpha$ denote the nominal simultaneous coverage probability, and let

$\hat{q}_+(1-\alpha/2)$ denote the empirical $(1-\alpha/2)$-quantile of a collection of bootstrap samples generated in the manner of $\mathcal{M}_n^*$. Also, let $\underline{\mathcal{M}}_n^*$ denote the counterpart of $\mathcal{M}_n^*$ that is obtained by replacing $\max_{1\leq j\leq p}$ with $\min_{1\leq j\leq p}$ in equation (1.2), and let $\hat{q}_-(\alpha/2)$ denote the empirical $(\alpha/2)$-quantile of a collection of bootstrap samples generated in the manner of $\underline{\mathcal{M}}_n^*$. In this notation, the confidence interval $\hat{\mathcal{I}}_j$ is defined by

$$\hat{\mathcal{I}}_j = \left\{ x \in \mathbb{R} \,:\, \hat{q}_-(\alpha/2) \leq \frac{1}{\sqrt{n}\hat{\sigma}_j^\tau} \sum_{i=1}^n \varphi_{\hat{t}_j}(X_{ij} - x) \leq \hat{q}_+(1-\alpha/2) \right\}. \tag{2.7}$$

Due to the fact that the functions $\varphi_{\hat{t}_1}(\cdot), \ldots, \varphi_{\hat{t}_p}(\cdot)$ are monotone, it is straightforward to compute all the endpoints of $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$.

To comment on the role of the partial standardization parameter $\tau \in [0,1]$, it provides a way to balance two opposing effects that occur in the extreme cases when τ is equal to 0 or 1. When $\tau = 0$, all of the intervals $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$ have the same width, which is clearly undesirable when the covariates fluctuate over different scales. Alternatively, when $\tau = 1$, all of the covariates will be on approximately "equal footing", which will tend to make the max statistic $\mathcal{M}_n$ sensitive to all p dimensions. This is undesirable in high-dimensional situations where the covariates fluctuate over different scales, because it eliminates a form of low-dimensional structure that can simplify the behavior of $\mathcal{M}_n$ when $\tau < 1$. To see this, consider a case where $\tau = 0$ and $\sigma_1, \ldots, \sigma_d$ are much larger than $\sigma_{d+1}, \ldots, \sigma_p$ for some $d \ll p$. In this case, the maximizing index for $\mathcal{M}_n$ is likely to reside in the small subset $\{1, \ldots, d\} \subset \{1, \ldots, p\}$. Thus, the behavior of $\mathcal{M}_n$ will be mainly governed by the first d covariates, which intuitively allows for more accurate distributional approximations of $\mathcal{M}_n$. Accordingly, as was originally proposed in (Lopes et al., 2020), it is natural to select an intermediate value of τ between 0 and 1 that can mitigate the unwanted effects

at $\tau \in \{0, 1\}$. In Section 4.2, we provide empirical results that support the recommended default of $\tau = 0.9$ and show that variations in the choice of τ affect the bootstrap in a very gradual way.

One more aspect of our method to discuss is computational cost. To generate a single bootstrap sample $\mathcal{M}_n^*$ requires $\mathcal{O}(np)$ flops, and hence the runtime for generating a collection of B bootstrap samples sequentially is $\mathcal{O}(Bnp)$. By comparison, the corresponding runtimes for the other robust approaches to bootstrapping max statistics discussed in the introduction are as follows: $\mathcal{O}(Bn^2p)$ for the method based on the Hodges-Lehmann estimator in (Fan et al., 2023), $\mathcal{O}(Bnp)$ for the self-normalization method adapted from (Liu and Shao, 2014) (to be described further in Section 4), and $\mathcal{O}(Bnp)$ for the truncation and trimming methods in (Kock and Preinerstorfer, 2025) (up to factors involving $\log(n)$ that may depend on the choice of sorting algorithms). In the case of the method from (Fan et al., 2023), the appearance of n^2 rather than n in the runtime is due to the fact that the Hodges-Lehmann estimator involves computing the median of a set of $\binom{n}{2}$ values. It should also be mentioned that all of the runtimes can be reduced by a factor of B if the methods are implemented to generate their bootstrap samples in parallel.

3. Theory

Our theoretical analysis is framed in terms of a sequence of models that are implicitly embedded in a triangular array whose rows are indexed by n. In this context, all model parameters are allowed to vary with n, except when stated otherwise. In particular, the dimension $p = p(n)$ is regarded as a function of n, and hence, if a quantity does not

depend on n, then it does not depend on p either.

To state our model assumptions, recall that the sorted coordinate-wise variances of X_1 are denoted as $\sigma^2_{(1)} \geq \cdots \geq \sigma^2_{(p)}$, and for any $d \in \{1, \ldots, p\}$, let $J(d)$ be a set of d indices in $\{1, \ldots, p\}$ that satisfies $\{\sigma_j^2 \,|\, j \in J(d)\} = \{\sigma^2_{(1)}, \ldots, \sigma^2_{(d)}\}$. In addition, let $R(d)$ denote the $d \times d$ correlation matrix associated with the covariates $\{X_{1j}\}_{j \in J(d)}$.

Assumption 1. The observations $X_1, \ldots, X_{n+m_n} \in \mathbb{R}^p$ are i.i.d., and there are constants $C \geq 1$, $\beta > 0$, and $\delta \geq \epsilon > 0$ not depending on n such that the following conditions hold:

(i) For all $v \in \mathbb{R}^p$, $\big\|\langle v, X_1 - \mathbf{E}(X_1)\rangle\big\|_{L^{4+\delta}} \leq C\big\|\langle v, X_1 - \mathbf{E}(X_1)\rangle\big\|_{L^2}$ holds.

(ii) For all $j = 1, \ldots, p$, the random variable X_{1j}/σ_j has a Lebesgue density f_j such that $\|f_j\|_{L^\infty} \leq C$.

(iii) For all $j = 1, \ldots, p$, the inequalities $\frac{1}{C}\sigma^2_{(1)} j^{-2\beta} \leq \sigma^2_{(j)} \leq C\sigma^2_{(1)} j^{-2\beta}$ hold.

(iv) If $l_n = \big\lceil n^{\frac{\epsilon}{24(\beta \vee 1)}} \wedge p \big\rceil$, then $\|R(l_n)\|_F^2 \leq C l_n^{2-\frac{1}{C}}$.

Remarks. All of the conditions in Assumption 1 are invariant to shifting $X_1 \mapsto X_1 + v$ for fixed $v \in \mathbb{R}^p$, and scaling $X_1 \mapsto cX_1$ for fixed $c \neq 0$. The following paragraphs provide several examples that address each of the conditions (i)-(iv). Also, it is straightforward to combine the examples to construct a wide assortment of data-generating distributions that satisfy all of the conditions in Assumption 1 simultaneously.

Examples of $L^{4+\delta}$-L^2 moment equivalence. In recent years, moment assumptions similar to condition (i) have been adopted in many analyses of robust statistical methods for high-dimensional data (Lugosi and Mendelson, 2019; Ke et al., 2019; Mendelson and Zhivotovskiy, 2020; Roy et al., 2021; Abdalla and Zhivotovskiy, 2022). As

shown in Proposition 1, condition (i) is compatible with the classes of *elliptical* and *separable* models (also known as independent component models), which are widely used in areas such as multivariate analysis, random matrix theory, and signal processing (Kotz et al., 2019; Bai and Silverstein, 2010; Comon and Jutten, 2010).

To be precise, we say that an observation X_1 with mean μ and covariance matrix Σ has an elliptical distribution if it can be represented as $X_1 = \mu + \eta_1 \Sigma^{1/2} U_1$, where $U_1 \in \mathbb{R}^p$ is uniformly distributed on the unit sphere, and η_1 is a non-negative scalar random variable that is independent of U_1 and normalized by $\mathbf{E}(\eta_1^2) = p$. On the other hand, we say that X_1 has a separable distribution if it can be represented as $X_1 = \mu + \Sigma^{1/2}\zeta_1$, where $\zeta_1 = (\zeta_{11}, \ldots, \zeta_{1p})$ has i.i.d. entries with $\mathbf{E}(\zeta_{11}) = 0$, and $\mathrm{var}(\zeta_{11}) = 1$.

Proposition 1. *Conditions (i) and (ii) hold simultaneously if one of the following two conditions holds for some $\delta > 0$ that is fixed with respect to n.*

(1) The observation X_1 is drawn from an elliptical distribution such that $\|\eta_1\|_{L^{4+\delta}} \lesssim \sqrt{p}$, and the random variable X_{11}/σ_1 has a Lebesgue density f_1 such that $\|f_1\|_{L^\infty} \lesssim 1$.

(2) The observation X_1 is drawn from a separable distribution with $\max_{1\le j\le p} \|\zeta_{1j}\|_{L^{4+\delta}} \lesssim 1$, and each random variable ζ_{1j} has a Lebesgue density g_j such that $\max_{1\le j\le p} \|g_j\|_{L^\infty} \lesssim 1$.

The proof is provided in Appendix G.

Examples of variance decay. In various settings, the sorted coordinate-wise variances $\sigma^2_{(1)} \ge \cdots \ge \sigma^2_{(p)}$ naturally exhibit a decay profile.

1. Principal components analysis. In the context of principal components analysis, it is common to assume that the sorted eigenvalues $\lambda_1(\Sigma) \ge \cdots \ge \lambda_p(\Sigma)$ of Σ satisfy

$\lambda_j(\Sigma) \lesssim j^{-\gamma}$ for some $\gamma > 0$, and it can be shown that this implies $\sigma^2_{(j)} \lesssim j^{-2\beta}$ for some other decay parameter $\beta > 0$ (Lopes et al., 2020, Proposition 2.1).

2. Mean-variance proportionality. Another scenario where variance decay arises is when the coordinate-wise means and variances are connected by a proportionality relationship of the form $\sigma_j^2 \propto |\mu_j|^\gamma$, for some fixed exponent $\gamma > 0$. This occurs within many sub-families of distributions, including Gamma, Weibull, inverse Gaussian, and Pareto. In applications that involve sparse modelling of high-dimensional mean vectors, a classical assumption is that the sorted coordinate-wise means have a decay profile (Johnstone, 2019), and thus, when such a proportionality relationship holds, it follows that variance decay must also occur.

3. Functional data analysis. One more set of examples is related to functional data analysis, where function-valued observations $\Psi_1, \ldots, \Psi_n$ in a Hilbert space are often studied through their projections under a finite number of orthonormal basis functions $\{\phi_j\}_{1 \le j \le p}$. That is, the ith projected observation has the form $X_i = (\langle \Psi_i, \phi_1 \rangle, \ldots, \langle \Psi_i, \phi_p \rangle) \in \mathbb{R}^p$. In connection with our work, the important point is that under standard assumptions in functional data analysis, it can be shown that the sorted coordinate-wise variances of X_i have a decay profile (Lopes et al., 2020). In fact, this occurs even when the random functions $\Psi_1, \ldots, \Psi_n$ have rough sample paths, which we illustrate empirically in Figure 2.

Examples of correlation matrices. To interpret the condition (iv), it should be noted that the inequality $\|R(l_n)\|_F^2 \le l_n^2$ always holds, since $\|A\|_F^2 \le \mathrm{tr}(A)^2$ holds for any positive semidefinite matrix A. So, in this sense, condition (iv) is quite mild, as C may be taken to be arbitrarily large. Moreover, the correlation structure of the variables indexed by

$\{1, \ldots, p\} \setminus J(l_n)$ is *completely unrestricted*. With regard to the constant 24 appearing in the definition of l_n, it has no special importance, and is used for theoretical convenience. Below, we describe several classes of $p \times p$ correlation matrices $R = R(p)$ for which the sub-matrix $R(l_n)$ satisfies condition (iv).

1. Decaying correlation functions. Let $\rho : [0, \infty) \to [0, 1]$ be any continuous convex function satisfying $\rho(0) = 1$, and $\rho(t) \leq ct^{-\gamma}$ for some fixed constants $c > 0, \gamma > 0$, and all $t \geq 0$. By Pólya's criterion (Pólya, 1949), a matrix whose ij entry is defined by $\rho(|i - j|)$ is a correlation matrix that satisfies condition (iv). Matrices of this type include many well-known examples, such as those of the autoregressive and banded types, e.g. $R_{ij} = r^{|i-j|}$ for some fixed $r \in (0, 1)$, and $R_{ij} = \max\left\{0, 1 - \frac{|i-j|}{b}\right\}$ for some fixed $b > 0$.

2. Diverging operator norm. If the operator norm of R satisfies $\|R\|_{\text{op}} \leq C l_n^{\frac{1}{2} - \frac{1}{2C}}$, then Assumption 1(iv) holds. This can be seen by noting that $\|R(l_n)\|_F^2 \leq l_n \|R(l_n)\|_{\text{op}}^2 \leq l_n \|R\|_{\text{op}}^2$. In particular, since l_n increases with n and p, this shows that condition (iv) can hold even when the operator norm of R diverges asymptotically.

3. Block structure. Suppose that R is formed by concatenating k correlation matrices along its diagonal, with sizes $\nu_1 \times \nu_1, \ldots, \nu_k \times \nu_k$, so that $\nu_1 + \cdots + \nu_k = p$. If the condition $\max\{\nu_1, \ldots, \nu_k\} \leq C l_n^{1 - \frac{1}{C}}$ holds, then so does condition (iv). This follows from the observation that no row of R can have a squared ℓ_2 norm larger than $\max\{\nu_1, \ldots, \nu_k\}$, and so $\|R(l_n)\|_F^2 \leq l_n \max\{\nu_1, \ldots, \nu_k\}$.

4. Convex combinations and permutations. If R and R' denote any correlation matrices corresponding to the previous examples, then for any $t \in [0, 1]$, the correlation matrix

$tR + (1-t)R'$ satisfies condition (iv). Furthermore, if Π is a $p \times p$ permutation matrix, then $\Pi R \Pi^\top$ is also a correlation matrix that satisfies condition (iv). These operations considerably extend the examples that have decaying correlation functions or block structure.

The following theorem is our main result.

Theorem 1. *Fix any constant $\tau \in [0,1)$ with respect to n, and suppose that Assumption 1 holds with the values of $\delta \geq \epsilon > 0$ stated there. In addition, suppose that the hold-out set consists of $m_n \asymp n$ observations that are partitioned into $b_n \asymp \log(n)$ blocks. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\sup_{s \in \mathbb{R}} \Big| \mathbf{P}(\mathcal{M}_n \leq s) - \mathbf{P}(\mathcal{M}_n^* \leq s | X) \Big| \;\leq\; c\, n^{-\frac{1}{2}+\epsilon}$$

occurs with probability at least $1 - c\, n^{-\delta/4}$.

Remarks. In addition to the fact that the rate of approximation is near $n^{-1/2}$ and is independent of the dimension p, it should also be noted that the decay parameter β is allowed to be arbitrarily small. Thus, the result demonstrates that the robust max statistic is able to reap substantial benefits from only a slight degree of variance decay structure.

In relation to previous work on ordinary max statistics such as $\max_{1 \leq j \leq p} \sqrt{n}(\bar{X}_j - \mu_j)$, the novelty in the proof of Theorem 1 is that it accounts for several effects that arise from heavy-tailed covariates and the robust design of the statistic $\mathcal{M}_n$. A particularly important example of such an effect is the bias that is introduced by the truncation functions $\varphi_{\hat{t}_j}$, and this is addressed in the proofs of Propositions 4 and 7, as well as

Lemmas 3, 4, and 7. Furthermore, our work accounts for the fluctuations of the robust MOM estimates $\tilde{X}_j$ and $\hat{\sigma}_j$ that are used to partially standardize $\mathcal{M}_n^*$, and this is done in Lemmas 12-15.

4. Numerical experiments

This section addresses the practical performance of the proposed method when applied to both Euclidean and functional data with heavy tails. We also include comparisons with the alternative methods mentioned in the introduction, as well as real-data examples involving heavy-tailed financial data.

4.1 Summary of alternative methods

Before describing the design of our experiments, we briefly summarize the alternative methods for comparison in the task of constructing simultaneous confidence intervals for the mean parameters $\mu_1, \ldots, \mu_p$. The methods include one based on the Hodges–Lehmann (HL) estimator proposed in (Fan et al., 2023), a self-normalization (SN)–based bootstrap method adapted from (Liu and Shao, 2014), as well as several methods based on trimming and truncation developed in (Kock and Preinerstorfer, 2025).

Hodges-Lehmann (HL). The HL pseudomedian estimator based on the jth covariates $X_{1j}, \ldots, X_{nj}$ is defined as $\hat{\theta}_j^{\mathrm{HL}} = \mathrm{median}\{\frac{1}{2}(X_{ij} + X_{lj}) : 1 \le i < l \le n\}$, i.e. the median of two-point averages taken over all $\binom{n}{2}$ pairs of distinct integers $i, l \in \{1, \ldots, n\}$ (Hodges and Lehmann, 1963). The paper (Fan et al., 2023) proposed the bootstrap version $\hat{\theta}_j^* = \mathrm{median}\{\frac{1}{2}(X_{ij} + X_{lj}) : i \neq l \in \mathcal{S}\}$, where the set $\mathcal{S}$ is drawn uniformly at random from all

subsets of $\{1,\dots,n\}$. For a desired nominal simultaneous coverage probability of $1-\alpha$, let $\hat{q}^{\mathrm{HL}}$ denote the empirical $(1-\alpha)$-quantile of a collection of bootstrap samples generated in the manner of $\max_{1\le j\le p}|\hat{\theta}_j^* - \hat{\theta}_j|$. In this notation, the paper (Fan et al., 2023) proposed simultaneous confidence intervals defined by $[\hat{\theta}_j \pm \hat{q}^{\mathrm{HL}}]$ for $j = 1,\dots,p$.

Self-normalization (SN). The SN method that we describe here does not seem to have been formally proposed in previous work, but it is an adaptation of an approach developed in (Liu and Shao, 2014) for a different task. For each $i = 1,\dots,n$ define the vector $\hat{X}_i \in \mathbb{R}^p$ in terms of the truncated covariates $\hat{X}_{ij} = X_{ij}\,1\{|X_{ij}| \le (n/\log p)^{1/6}\}$, for $j = 1,\dots,p$. Next, let $\hat{X}_1^*,\dots,\hat{X}_n^*$ be sampled with replacement from $\hat{X}_1,\dots,\hat{X}_n$, and define $\hat{t}_j^*$ to be the Student's t statistic computed from the values $\hat{X}_{1j}^* - \hat{X}_j,\dots,\hat{X}_{nj}^* - \hat{X}_j$, where $\hat{X}_j = \frac{1}{n}\sum_{i=1}^n \hat{X}_{ij}$. Let $\hat{q}_+^{\mathrm{SN}}$ denote the empirical $(1-\alpha/2)$-quantile of a collection of bootstrap samples generated in the manner of $\max_{1\le j\le p}\hat{t}_j^*$. Similarly, let $\hat{q}_-^{\mathrm{SN}}$ denote the $(\alpha/2)$-quantile of a collection of corresponding bootstrap samples with $\min_{1\le j\le p}$ replacing $\max_{1\le j\le p}$. The SN intervals are defined by $[\bar{X}_j \pm \frac{1}{\sqrt{n}}\tilde{\sigma}_j\hat{q}_\mp^{\mathrm{SN}}]$ for $j = 1,\dots,p$, where $\bar{X}_j$ and $\tilde{\sigma}_j$ are the usual sample mean and standard deviation of $X_{1j},\dots,X_{nj}$.

Trimming and truncation. The paper (Kock and Preinerstorfer, 2025) proposed several methods for constructing simultaneous confidence intervals, based on various combinations of trimming, truncation, and standardization. When these methods were implemented with the suggested tuning parameters in the settings described below in Section 4.2, we observed that for nominal simultaneous coverage probabilities of 90% or 95%, the actual simultaneous coverage probabilities were often below 70%. Since these results are not competitive with the other three methods to be compared later, and since the

definition of the trimming and truncation methods are intricate, we do not discuss them in further detail.

4.2 Euclidean data

Here, we consider the task of constructing simultaneous confidence intervals for the entries of the mean vector $(\mu_1, \ldots, \mu_p) = \mathbf{E}(X_1)$, based on i.i.d. observations in $\mathbb{R}^p$.

Design of experiments. For each pair $(n + m_n, p)$ in the set $\{500\} \times \{100, 500, 1000\}$, and each of the data-generating distributions described below, we performed 500 Monte Carlo trials. For simplicity, the mean vector $(\mu_1, \ldots, \mu_p)$ was always chosen to be the zero vector, which is without loss of generality for methods that are shift invariant. In all trials, the proposed confidence intervals $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$ were constructed according to (2.7), using 500 bootstrap samples, and using choices of 90% and 95% for the nominal simultaneous coverage probability $1-\alpha$. Also, the proposed method was always applied with truncation parameters set to $\hat{t}_j = \sqrt{n}\hat{\sigma}_j$ for all $j = 1, \ldots, p$, and with the partial standardization parameter set to $\tau = 0.9$. Lastly, in all trials, the number of hold-out observations was set to $m_n = 50$, and the block length for the MOM estimates was set to $\ell_n = 10$. (The HL and SN methods were implemented based on a total sample size of $n + m_n$.)

Data-generating distributions. The data were generated in four ways, based on an elliptical distribution and a separable distribution with two choices of covariance matrices.

Elliptical distribution. The elliptical observations have the form $X_1 = \mu + \eta_1 \Sigma^{1/2} U_1$, where $U_1 \in \mathbb{R}^p$ is uniformly distributed on the unit sphere, and $\eta_1 \geq 0$ is a random variable that is independent of U_1 such that $3\eta_1^2/(2p)$ follows an F distribution with p and 6 degrees of freedom. This distribution for X_1 is commonly known as a *multivariate t-distribution on*

6 degrees of freedom (Muirhead, 2009).

Separable distribution. The separable observations have the form $X_1 = \mu + \Sigma^{1/2}\zeta_1$, where $\zeta_1 = (\zeta_{11}, \ldots, \zeta_{1p})$ has i.i.d. entries that are standardized Pareto random variables. Specifically, $\zeta_{11} = (\omega_{11} - \mathbf{E}(\omega_{11}))/\sqrt{\text{var}(\omega_{11})}$, where ω_{11} is drawn from a Pareto distribution whose density is given by $x \mapsto 6x^{-7}1\{x \geq 1\}$.

Covariance matrices. For both the elliptical and separable distributions, we constructed the covariance matrix of X_1 in the form $\Sigma = D^{1/2}RD^{1/2}$, where $D = \text{diag}(\text{var}(X_{11}), \ldots, \text{var}(X_{1p}))$, and R is the correlation matrix of X_1. The correlation matrix was chosen to be one of the following two types

$$R_{ij} = \begin{cases} 0.5^{|i-j|} & \text{(autoregressive)} \\ 1\{i=j\} + \frac{1\{i \neq j\}}{4(i-j)^2} & \text{(algebraic decay).} \end{cases}$$

In all cases, the matrix D was chosen to satisfy $D_{jj}^{1/2} = j^{-1/2}$ for all $j = 1, \ldots, p$, ensuring that the entries of X_1 have variance decay.

Discussion of results. In Table 1, we report empirical simultaneous coverage probabilities and width measures, for the proposed method (denoted PM), as well as the HL and SN methods. The simultaneous coverage probabilities were computed as the fraction of the 500 Monte Carlo trials in which all p intervals of a given method contained the corresponding parameters $\mu_1, \ldots, \mu_p$. The width measure was computed as the median width of the p intervals, averaged over the 500 trials, and it is shown in parentheses below the simultaneous coverage probabilities.

Across all of the settings, PM and SN deliver simultaneous coverage probabilities that are within about 1 or 2 percent of the nominal level, demonstrating that they are reliably

Table 1: Comparison of simultaneous coverage probability and confidence interval width, where PM refers to the proposed method. The nominal simultaneous coverage probability is either 90% or 95%, as indicated by the third column.

R	Distribution	$1-\alpha$	$p=100$			$p=500$			$p=1000$		
			PM	HL	SN	PM	HL	SN	PM	HL	SN
auto-regressive	elliptical	0.95	0.942	0.942	0.954	0.956	0.938	0.960	0.962	0.962	0.942
			(0.057)	(0.165)	(0.044)	(0.031)	(0.165)	(0.022)	(0.023)	(0.166)	(0.016)
		0.90	0.914	0.892	0.902	0.918	0.900	0.916	0.908	0.916	0.890
			(0.053)	(0.142)	(0.041)	(0.029)	(0.142)	(0.021)	(0.022)	(0.142)	(0.016)
	separable	0.95	0.956	0.006	0.952	0.946	0.002	0.934	0.954	0	0.946
			(0.065)	(0.143)	(0.045)	(0.035)	(0.144)	(0.023)	(0.027)	(0.143)	(0.017)
		0.90	0.922	0	0.906	0.904	0.002	0.898	0.910	0	0.904
			(0.059)	(0.123)	(0.042)	(0.033)	(0.124)	(0.022)	(0.025)	(0.123)	(0.016)
algebraic decay	elliptical	0.95	0.946	0.938	0.946	0.942	0.942	0.95	0.958	0.962	0.956
			(0.058)	(0.167)	(0.044)	(0.031)	(0.166)	(0.022)	(0.023)	(0.166)	(0.016)
		0.90	0.914	0.890	0.912	0.910	0.898	0.898	0.916	0.904	0.910
			(0.053)	(0.144)	(0.041)	(0.029)	(0.143)	(0.021)	(0.022)	(0.144)	(0.016)
	separable	0.95	0.968	0	0.954	0.954	0	0.948	0.958	0	0.968
			(0.067)	(0.141)	(0.046)	(0.037)	(0.142)	(0.023)	(0.028)	(0.141)	(0.017)
		0.90	0.922	0	0.906	0.910	0	0.910	0.912	0	0.916
			(0.062)	(0.122)	(0.043)	(0.034)	(0.123)	(0.022)	(0.026)	(0.122)	(0.016)

calibrated. On the other hand, HL is only well calibrated in the cases of elliptical models, where the covariates have distributions that are symmetric around 0. In the cases of separable models where the covariates have distributions that are *not* symmetric around 0, the simultaneous coverage probabilities of HL are far from the nominal level. This can be explained by the fact that HL is designed for simultaneous inference on the coordinate-wise pseudomedians—which may differ from the coordinate-wise means if the covariates have asymmetric distributions. By contrast, PM and SN have a scope of application for inference on coordinate-wise means that is not limited by asymmetric distributions.

With regard to the width measure, Table 1 shows that the intervals produced by PM and SN are much tighter than those produced by HL, while the width advantage of SN over PM is relatively small. The larger widths of the HL intervals are to be expected, because all of these widths must be the same across all p coordinates, whereas the PM and SN intervals are adaptive to the scale of each covariate. This difference between the methods is also reflected in another pattern—which is that the width measure for PM and SN decreases as p increases, whereas the width measure for HL stays essentially constant as p increases, since HL uses unstandardized covariates.

Influence of τ. To illustrate how the choice of τ affects the performance of PM, we repeated the experiments from Table 1 with the values $\tau \in \{0.6, 0.7, 0.8, 0.9\}$, and the results are shown below in Table 2, with coverage probabilities and interval widths reported in the same format. The results show that the choice of τ affects the bootstrap method in a mild way, with the coverage probability being mostly unaffected by τ, and the width decreasing slightly as τ increases. Another point to mention is that the set

Table 2: Comparison of simultaneous coverage probability and confidence interval width (parentheses) across different τ values for the proposed method. In all cases, the nominal simultaneous coverage probability is 95%.

R	Distribution	τ	$p = 100$	$p = 500$	$p = 1000$
auto-regressive	elliptical	0.6	0.932 (0.068)	0.950 (0.043)	0.940 (0.035)
		0.7	0.944 (0.062)	0.948 (0.035)	0.956 (0.028)
		0.8	0.942 (0.058)	0.956 (0.032)	0.952 (0.024)
		0.9	0.938 (0.058)	0.960 (0.031)	0.958 (0.023)
	separable	0.6	0.960 (0.073)	0.954 (0.046)	0.952 (0.037)
		0.7	0.960 (0.067)	0.952 (0.039)	0.958 (0.030)
		0.8	0.960 (0.065)	0.942 (0.036)	0.960 (0.027)
		0.9	0.958 (0.065)	0.954 (0.035)	0.954 (0.027)
algebraic decay	elliptical	0.6	0.942 (0.069)	0.946 (0.043)	0.940 (0.035)
		0.7	0.944 (0.062)	0.948 (0.036)	0.946 (0.028)
		0.8	0.940 (0.059)	0.948 (0.032)	0.946 (0.024)
		0.9	0.946 (0.058)	0.942 (0.031)	0.958 (0.023)
	separable	0.6	0.964 (0.074)	0.966 (0.046)	0.966 (0.037)
		0.7	0.960 (0.069)	0.956 (0.040)	0.954 (0.031)
		0.8	0.960 (0.067)	0.952 (0.037)	0.966 (0.028)
		0.9	0.962 (0.067)	0.956 (0.037)	0.956 (0.028)

$\{0.6, 0.7, 0.8, 0.9\}$ contains values of τ that are likely to be relevant in practice. Indeed, recall that in the extreme case of $\tau = 0$, all of the intervals $\widehat{\mathcal{I}}_1, \ldots, \widehat{\mathcal{I}}_p$ have the same width, which is undesirable when the covariates fluctuate over different scales. In other words, the choice of $\tau = 0$ leads to intervals that are too wide for the covariates with small fluctuations. This issue also occurs for small positive values of τ. At the other extreme, if full standardization is used with $\tau = 1$, then the variance decay structure is eliminated, which motivates using values of τ that are somewhat less than 1. In summary, the table supports the recommended default choice of $\tau = 0.9$, since this value tends to produce the smallest median width among $\{0.6, 0.7, 0.8, 0.9\}$ while maintaining a simultaneous coverage probability close to 95%.

4.3 Functional data

Beyond heavy-tailed Euclidean data, our approach to inference with robust max statistics can be applied to heavy-tailed functional data. Below, we study the problem of detecting a non-zero drift in functional observations that arise from geometric Brownian motion (GBM). This is a heavy-tailed stochastic process in the sense that its marginals are lognormal, which is a standard example of heavy-tailed univariate distribution (Foss et al., 2011; Nair et al., 2022). The sample paths of GBM present additional challenges from the standpoint of functional data analysis, because they are rough. Furthermore, GBM is of broad interest in financial applications, where it is widely used for modelling securities prices (Oksendal, 2013).

Problem formulation. A sample path of GBM on the unit interval has the form $t \mapsto \exp\left((\mu_0 - \varsigma_0^2/2)t + \varsigma_0 W(t)\right)$ for $t \in [0, 1]$, where $W(t)$ is a standard Brownian motion,

$\mu_0 \in \mathbb{R}$ is the drift parameter, and $\varsigma_0^2 \geq 0$ is the volatility parameter. To formalize the detection of non-zero drift in a way that allows us to incorporate natural alternative hypotheses, we will allow for more general sample paths $S(t)$ of the form

$$S(t) = \exp\big((h\mu(t) - \varsigma_0^2/2)t + \varsigma_0 W(t)\big), \tag{4.8}$$

where $\mu(t)$ is a fixed real-valued function on $[0,1]$ such that $S(t)$ resides in $L^2[0,1]$ almost surely, and $h \geq 0$ is a fixed parameter that measures the "distance" from the null hypothesis of zero drift that occurs when $h = 0$. The parameters $\mu(t)$ and ς_0 are treated as unknown. Under these conditions, we are interested in using a dataset $S_1(t), \ldots, S_{n+m_n}(t)$ of i.i.d. samples of $S(t)$ to address the hypothesis testing problem

$$\mathsf{H}_0 : h = 0 \quad \text{vs.} \quad \mathsf{H}_1 : h > 0. \tag{4.9}$$

In particular, different choices of the function $\mu(t)$ correspond to different alternatives, and later on, we will present numerical results for several choices.

Testing procedure. The sample path formula (4.8) implies that $\mathbf{E}(S(t)) = \exp(h\mu(t)t)$ for all $t \in [0,1]$. For this reason, our procedure will seek to detect whether or not the function $\mathbf{E}(S(t)) - 1$ is identically 0. This will be done by expanding $\mathbf{E}(S(t)) - 1$ in the form $\sum_{j=1}^{\infty} \beta_j \phi_j(t)$, where $\{\phi_j(t)\}_{j\geq 1}$ is the Fourier cosine basis for $L^2[0,1]$. To proceed, note that $\mathbf{E}(S(t)) - 1$ is equal to the zero function in the $L^2[0,1]$ sense if and only if $\beta_j = 0$ for all $j \geq 1$. This motivates a procedure based on testing the simultaneous hypotheses

$$\mathsf{H}_{0,j} : \beta_j = 0 \quad \text{for} \quad j = 1, \ldots, p, \tag{4.10}$$

where p is an integer large enough so that the coefficients $\beta_{p+1}, \beta_{p+2}, \ldots,$ are negligible for practical purposes. In particular, H_0 implies that $\mathsf{H}_{0,1}, \ldots, \mathsf{H}_{0,p}$ hold simultaneously,

and so a procedure that controls the simultaneous type I error rate for these hypotheses leads to one that controls the ordinary type I error rate for H_0.

If we define $X_i \in \mathbb{R}^p$ to contain the first p coefficients of $S_i(t) - 1$ with respect to $\{\phi_j(t)\}_{j\geq 1}$, then it follows that $\mathbf{E}(X_i) = (\beta_1, \ldots, \beta_p)$ for every $i = 1, \ldots, n + m_n$. Hence, we may apply our proposed method from Section 2 to $X_1, \ldots, X_{n+m_n}$ in order to construct confidence intervals $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$ for $\beta_1, \ldots, \beta_p$ with a nominal simultaneous coverage probability of $1 - \alpha$. Altogether, this means that if we reject H_0 when any of these intervals exclude 0, then this rejection rule corresponds to a test with a nominal level of at most α.

Design of experiments. To construct four natural choices of the pair $(\mu(t), \varsigma_0)$, we used historical data for the stocks of Apple, Nvidia, Moderna, and JPMorgan. With regard to the choice of ς_0^2, note that if a stock price is modeled as a realization of $S(t)$, then the pointwise variance of the cumulative log return is $\mathrm{var}((h\mu(t) - \varsigma_0^2/2)t + \varsigma_0 W(t)) = \varsigma_0^2 t$, and the time average of this quantity over the unit interval is $\varsigma_0^2/2$. (See (Ruppert and Matteson, 2011) for additional background.) Accordingly, for each of the four stocks, we selected ς_0 such that $\varsigma_0^2/2 = \int_0^1 s^2(t)\,\mathrm{d}t$, where $s^2(t)$ is the sample pointwise variance of the cumulative log return of the stock over many disjoint periods of unit length. (The data were prepared in the same manner as described in Section 4.4 below, except that a period of unit length here was taken to be 100 minutes.) Next, the selection of $\mu(t)$ was based on the fact that if a stock price is modeled with $S(t)$, then the pointwise expected cumulative log return is $(h\mu(t) - \varsigma_0^2/2)t$. Therefore, for a given value of $h = 1$ and ς_0, we defined $\mu(t)$ to satisfy $(\mu(t) - \varsigma_0^2/2)t = \bar{R}(t)$, where $\bar{R}(t)$ is the sample average of the

cumulative log return curves of a given stock over many disjoint periods of unit length.

For each of the four choices of $(\mu(t), \varsigma_0)$, and each value of h in an equispaced grid, the following procedure was repeated in 500 Monte Carlo trials. We generated i.i.d. realizations $S_1(t), \ldots, S_{n+m_n}(t)$ of the sample path defined in (4.8) with $n + m_n = 300$ and $m_n = 30$. For each of these sets of functional observations, we constructed simultaneous $(1 - \alpha)$-confidence intervals $\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p$ for the parameters $\beta_1, \ldots, \beta_p$, as described above, with $p = 100$ and $\alpha = 5\%$. Whenever any of these intervals excluded 0, a rejection was recorded, and the rejection rate among the 500 trials was plotted as a function of h in Figures 3a-3d. The corresponding rejection rates based on the simultaneous confidence intervals constructed from the HL and SN methods were also plotted in the same way. In all four figures, the nominal level of $\alpha = 5\%$ is marked with a dashed horizontal line.

To illustrate the characteristics of the simulated functional data, ten realizations of $S(t)$ based on $h = 0$ with ς_0 corresponding to Apple stock are plotted in Figure 1. In the same setting, Figure 2 displays estimates of the sorted values $\sigma_{(1)} \geq \cdots \geq \sigma_{(p)}$, where σ_j^2 is the variance of the jth Fourier coefficient of $S_1(t)$. In particular, Figure 2 shows a clear variance decay profile.

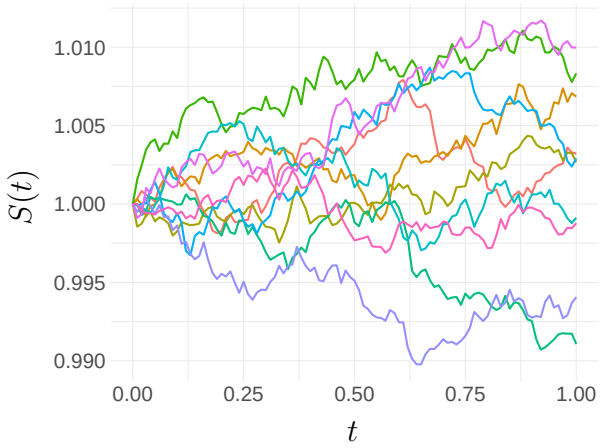


Figure 1: Representative sample paths of $S(t) = \exp\big((h\mu(t) - \varsigma_0^2/2)t + \varsigma_0 W(t)\big)$ when $h = 0$, and ς_0 is selected based on historical price data for Apple stock.

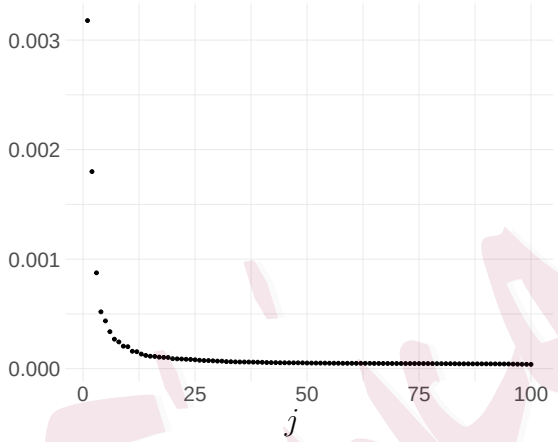
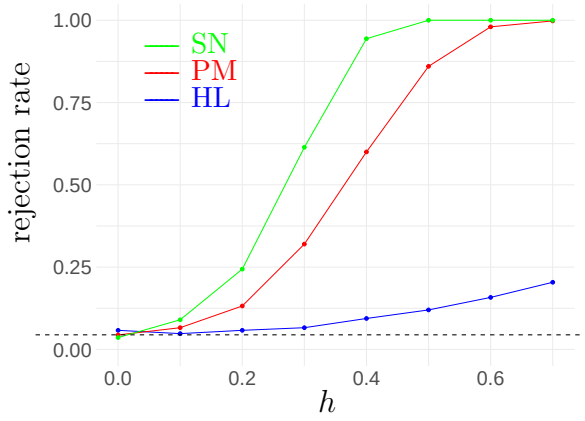


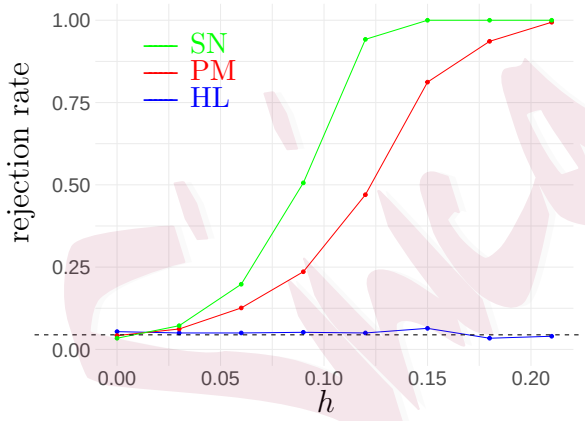
Figure 2: Estimates of the ordered values $\sigma_{(1)} \geq \cdots \geq \sigma_{(p)}$, where σ_j^2 denotes the variance of the jth Fourier coefficient of sample paths generated as in Figure 1.

Discussion of results. Figures 3a–3d display the rejection rate curves for the three methods PM, HL, and SN. Recall that $h = 0$ holds under the null hypothesis H_0, and so the value of a curve at $h = 0$ represents the empirical level. It is clear that in all four panels of Figure 3, the empirical levels of all methods closely match the nominal level of $\alpha = 5\%$, marked with a dashed horizontal line.

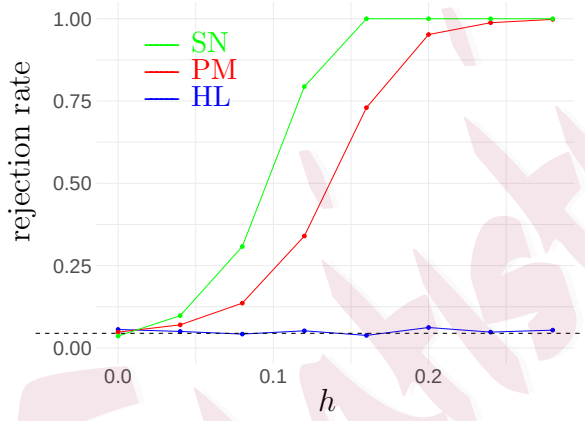
However, the methods differ in terms of power, which is represented by the values of the curves at $h > 0$. Overall, PM and SN have substantially higher power than HL, while the power of SN is somewhat higher than that of PM. The power advantage of PM and SN in comparison to HL is understandable in light of the results of Section 4.2, which showed that the confidence intervals produced by PM and SN were often much tighter than those of HL. (In the current context, tighter intervals allow 0 to be excluded more



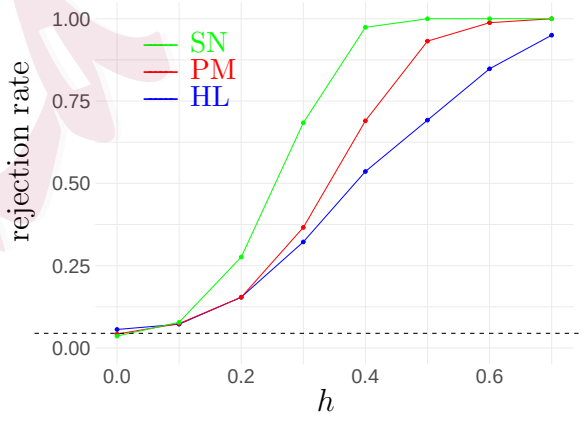
(a) Apple stock



(b) Nvidia stock



(c) JP Morgan stock



(d) Moderna stock

Figure 3: The panels compare the rejection rates of the methods PM, HL and SN when functional observations are generated in the form (4.8) and the parameters $(\mu(t), \varsigma_0)$ are based on historical stock price data for Apple, Nvidia, JP Morgan, and Moderna.

often, resulting in higher power.)

4.4 Illustration with financial data

In this subsection, we study an application that uses historical log return curves in order to screen candidate pairs of stocks for a strategy known as *pairs trading*. In addition to its interest from the viewpoints of finance and functional data analysis, this application illustrates two effects that are related to our proposed methodology: heavy tails and variance decay.

Background on pairs trading. Pairs trading is a well-established strategy that takes advantage of highly correlated pairs of stocks when their relative performance deviates from typical behavior (Vidyamurthy, 2004). When these situations occur, the strategy involves purchasing the stock that is underperforming, and selling short the stock that is overperforming—with the idea being that both trades will become profitable when the relative performance reverts to its mean. One of the main reasons for interest in pairs trading is that even though the individual performances of the stocks may be hard to predict, the relative performance may still be rather stable, which enhances the reliability of the strategy.

Below, we apply the method proposed in Section 2 to study a basic question that is essential to any pairs trading strategy. Starting from a collection of candidate pairs of stocks, we seek to identify which pairs have relative performance that is *unstable*. In other words, we seek to identify which pairs should be discarded from further consideration, because they are not well suited to pairs trading. Hence, this problem can be viewed as part of a preliminary filtering process that is done before any trading takes place.

Problem formulation. For a given pair of stocks, we measure their relative performance in terms of the difference $D(t)$ of their log returns (Ruppert and Matteson, 2011), which is viewed as a random function of time $t \in [0, 1]$. As a way of preparing an approximate sample of realizations of $D(t)$, we used historical stock data, as discussed later in this subsection. To formalize the question of whether the relative performance is stable or unstable, we consider the hypothesis testing problem

$$\mathsf{H}_0 : \text{The function } \mathbf{E}(D(t)) \text{ is constant. vs.}$$

$$\mathsf{H}_1 : \text{The function } \mathbf{E}(D(t)) \text{ is non-constant.}$$

Thus, if H_0 cannot be rejected, then the stocks underlying $D(t)$ are plausible candidates for pairs trading. It is also worth noting that a variety of other approaches have been proposed for selecting trading pairs (Vidyamurthy, 2004, Ch. 5-6), and the formulation above is only intended to serve as an illustration of our proposed methodology. More generally, the problem of testing whether or not a mean function is constant is of fundamental interest in other applications involving functional data.

Although the null hypothesis H_0 might seem easy to reject at first sight, it should be emphasized that H_0 is often difficult to reject for three reasons. First, the value of $\mathbf{E}(D(0))$ under H_0 is *unknown and unrestricted*, which means that H_0 allows for many possible values of $\mathbf{E}(D(0))$. Consequently, it is more difficult to reject H_0 than a simpler condition such as '$\mathbf{E}(D(t)) = c_0$ for all $t \in [0, 1]$', where c_0 is a particular value that is specified by the user. Second, as is commonly done in pairs trading, we used data for pairs of stocks involving similar businesses. Hence, this predisposes $\mathbf{E}(D(t))$ to be more stable than it would be for an arbitrarily chosen pair of stocks. Third, the log return of a

generic stock typically fluctuates near 0, which reflects the inherent difficulty of making profitable trades. Likewise, the difference of log returns $D(t)$ also tends to fluctuate near 0, making it difficult to reject H_0.

Data preparation. Price data for 10 pairs of stocks (listed in Table 3) were collected from the Alpha Vantage database (Alpha Vantage, 2024) during every minute of trading between March 1, 2024 and March 22, 2024, including pre-market and after-hours trading. The data were then divided into 30-minute intervals (normalized to unit length), and each interval was divided into time points $t_0, t_1, \ldots, t_{30}$ spaced one minute apart. Letting $P_1(t_j)$ and $P_2(t_j)$ denote the prices for a pair of stocks at t_j, the difference of log returns was computed as $D(t_j) = \log(P_1(t_j)/P_1(t_{j-1})) - \log(P_2(t_j)/P_2(t_{j-1}))$ for each $j = 1, \ldots, 30$. In this way, we obtained one discretely observed realization of the function $D(t)$ over each 30-minute interval. To promote independence among these functional observations, we only retained them from *every other* 30-minute interval, ensuring that they are separated by gaps of 30 minutes. We also excluded functional observations that were obtained from intervals with missing data. The resulting set of functional observations are denoted $D_1(t), \ldots, D_{n+m_n}(t)$, and the values of n and m_n are shown in Table 3.

Testing procedure. To test whether or not the function $\mathbf{E}(D(t))$ is constant with respect to $t \in [0, 1]$, we consider an expansion of the form $\mathbf{E}(D(t)) = \sum_{j=1}^{\infty} \gamma_j \phi_j(t)$, where $\{\phi_j(t)\}_{j \geq 1}$ is again the Fourier cosine basis for $L^2[0, 1]$. A convenient property of this basis is that H_0 is equivalent to the condition that $\gamma_j = 0$ for $j \geq 2$. (Note that $\phi_1(t)$ is a constant function, and so the first Fourier coefficient γ_1 may still be non-zero under H_0.) Similarly to our approach in Section 4.3, this leads us to consider the simultaneous

hypotheses

$$\mathsf{H}_{0,j} : \gamma_j = 0 \quad \text{for} \quad j = 2, \ldots, p+1, \tag{4.11}$$

where p is chosen large enough such that $\gamma_{p+2}, \gamma_{p+3}, \ldots$ are effectively negligible. As in Section 4.3, a procedure that controls the simultaneous type I error rate for $\mathsf{H}_{0,2}, \ldots, \mathsf{H}_{0,p+1}$ gives rise to one that controls the ordinary type I error rate for H_0. Despite this similarity with Section 4.3, there is still a notable difference, since the first Fourier coefficient is irrelevant in the current context. This is because Section 4.3 dealt with testing the null that a function is identically zero, whereas here we are testing the null that a function is constant.

We applied the methodology from Section 2 to the random vectors $X_1, \ldots, X_{n+m_n} \in \mathbb{R}^p$, with $p = 50$, where X_i is defined to contain the coefficients of $D_i(t)$ with respect to $\phi_2(t), \ldots, \phi_{p+1}(t)$, so that $\mathbf{E}(X_i) = (\gamma_2, \ldots, \gamma_{p+1})$. Letting $\hat{\mathcal{I}}_2, \ldots, \hat{\mathcal{I}}_{p+1}$ denote the resulting confidence intervals for $\gamma_2, \ldots, \gamma_{p+1}$ with nominal simultaneous coverage probability of $1-\alpha$, we reject H_0 with a nominal level of α whenever any of these intervals exclude 0.

Discussion of results. After testing 10 instances of H_0 based on 10 stock pairs, we recorded the resulting p-values in the column of Table 3 labeled PM. The column labeled HL reports the p-values that were computed by using the same testing strategy described above, but with simultaneous confidence intervals computed using the HL method, and likewise for the p-values in the column labeled SN. Looking at all 10 stock pairs as a whole, the table indicates that there is no clear winner among PM, HL, and SN, as none of the methods consistently produce smaller p-values than the others. Also, nearly all of the p-values are larger than commonly used significance thresholds, which may be viewed

as a reflection of the challenging aspects of the testing problem that were discussed earlier.

Table 3: Comparison of p-values produced by PM, HL, and SN

Company 1	Company 2	PM	HL	SN	n	m_n
Apple	Microsoft	0.252	0.318	0.044	230	20
Google	Meta	0.534	0.290	1	210	20
Tesla	General Motors	0.452	0.682	0.612	150	10
Amazon	Walmart	0.116	0.296	0.576	180	10
Pfizer	Moderna	1	0.886	1	150	10
Netflix	Disney	0.266	0.384	0.752	180	20
Nvidia	AMD	0.912	0.622	0.364	230	20
JPMorgan Chase	Bank of America	0.668	0.282	0.632	140	10
Shopify	eBay	0.404	0.270	0.476	100	10
UPS	FedEx	0.874	0.704	0.952	100	10

As an illustration of the difficulty of rejecting H_0, Figures 4a-4b display representative log return curves over 100-minute periods for the individual stocks in the (Apple, Microsoft) and (Google, Meta) pairs. Recalling that $D(t)$ is difference of the expected log return curves for each pair, these figures show that H_0 is challenging to reject because the individual log return curves have few sustained movements away from 0.

To give additional context to these results, it is worth commenting on some of the empirical characteristics of the random vectors $X_1, \ldots, X_{n+m_n}$ of Fourier coefficients. As was discussed in Section 4.2, the HL method is appropriate for inference on the mean pa-

rameters $(\gamma_2, \ldots, \gamma_{p+1}) = \mathbf{E}(X_1)$ when the coordinates of X_1 have symmetric distributions around 0. In Figures 5a-5d, we plot some representative histograms of these coordinate distributions for the pairs (Apple, Microsoft) and (Google, Meta), confirming that they are nearly symmetric. Hence, this indicates that the comparison among PM, HL and SN is fair. Another relevant aspect of the coordinate distributions is the heaviness of their tails (since all three methods are intended to be robust). On this point, Figures 6a-6d display normal QQ plots for the coordinate distributions. These QQ plots show clear departures from a linear trend, indicating that the tails of the coordinate distributions are indeed heavy. Lastly, since the PM method is designed to take advantage of variance decay among the Fourier coefficients, the presence of this decay is confirmed Figures 7a-7b.

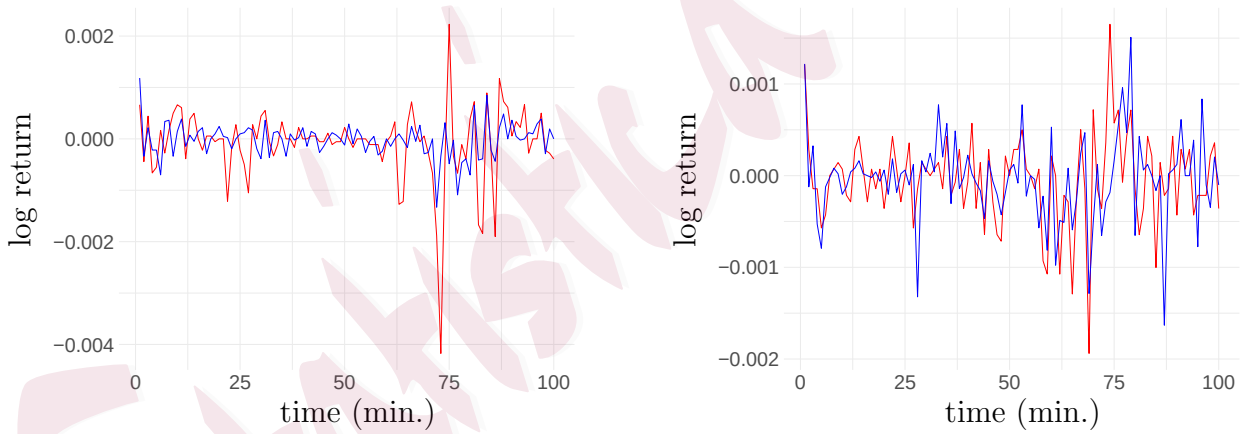


Figure 4: Panel (a) shows log return curves for Apple (red) and Microsoft (blue), while panel (b) shows log return curves for Google (red) and Meta (blue).

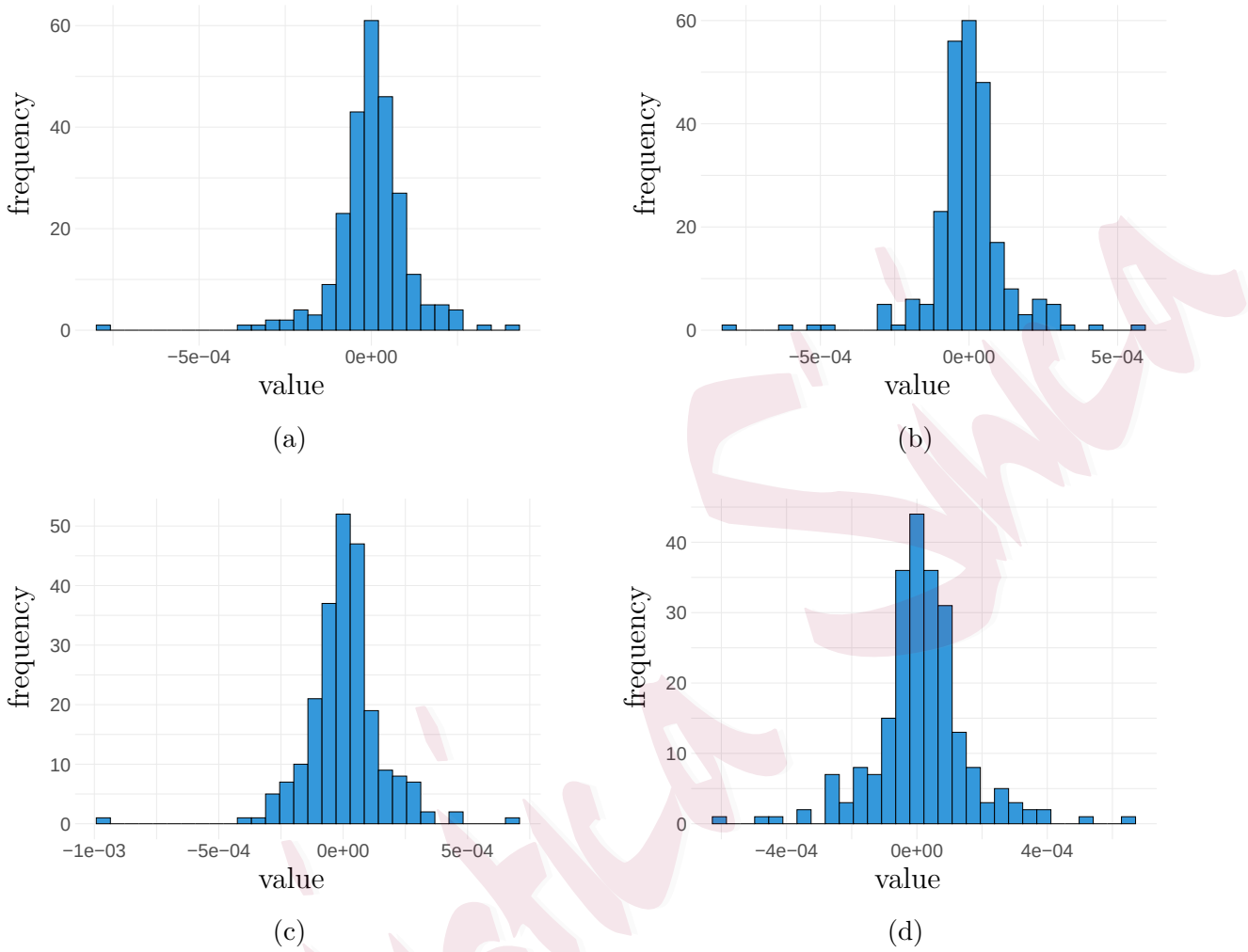


Figure 5: Panels (a) and (b) respectively show histograms of second and third Fourier coefficients of $D(t)$ for (Apple, Microsoft), while panels (c) and (d) show the corresponding histograms for (Google, Meta).

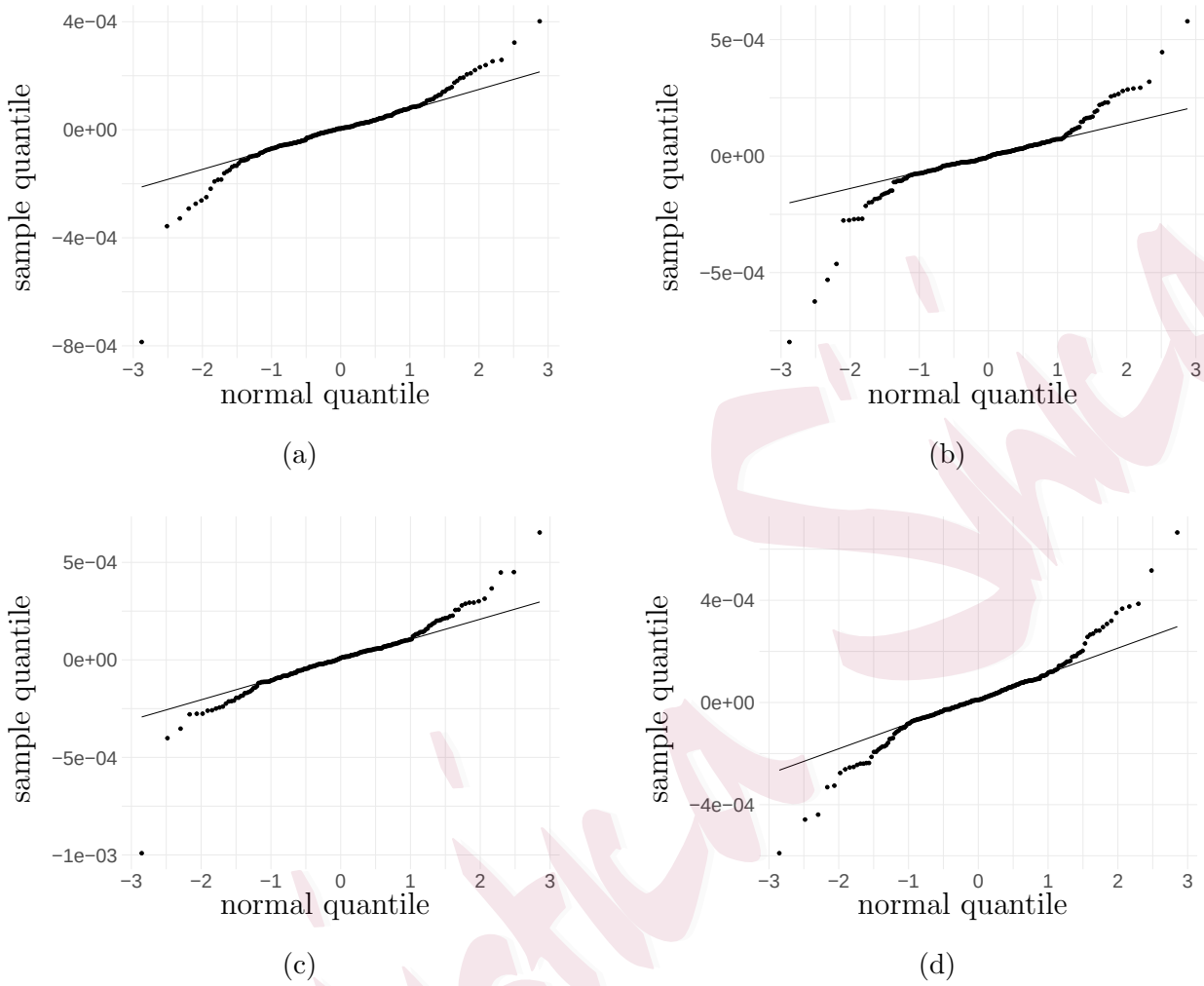
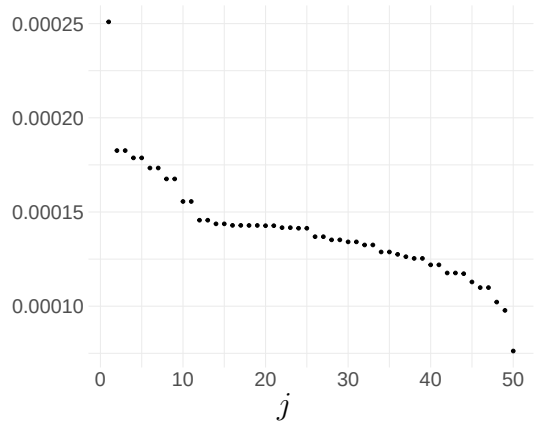
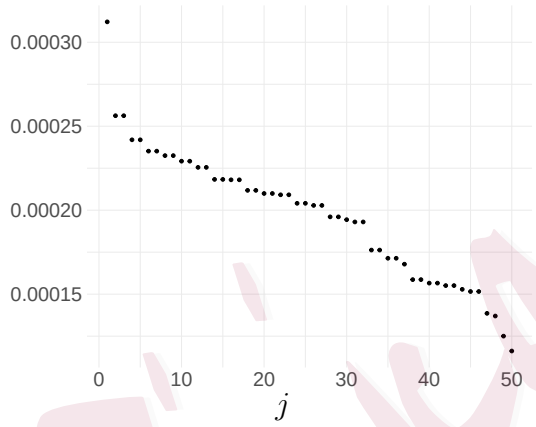


Figure 6: Panels (a) and (b) respectively show normal Q-Q plots of the second and third Fourier coefficient of $D(t)$ for (Apple, Microsoft), while panels (c) and (d) show the corresponding normal Q-Q plots for (Google, Meta)



(a)



(b)

Figure 7: Panel (a) shows the estimates of the ordered values $\sigma_{(1)} \geq \cdots \geq \sigma_{(p)}$, where σ_j^2 denotes the variance of the $(j+1)$th Fourier coefficient of $D(t)$ for (Apple, Microsoft). Panel (b) shows the corresponding estimates for (Google, Meta).

5. Conclusion

In this paper, we have proposed an approach to high-dimensional inference with heavy-tailed data that is based on bootstrapping *robust max statistics*. These statistics enhance the robustness of conventional max statistics using truncation and median-of-means estimates, and consequently, they have a scope of application that extends beyond light-tailed settings that much of the existing work on max statistics has focused on up to now. Under the assumption that the tails of the data are only constrained in terms of low-order moments, and that the covariates satisfy a natural structural condition known as variance decay, we have shown that the distributions of robust max statistics can be approximated via bootstrapping at a near $n^{-1/2}$ rate in the Kolmogorov metric, regardless of the data dimension p.

To address empirical performance, we have studied our method using both Euclidean and functional data, and have performed comparisons with a handful of other recently proposed methods for bootstrapping max statistics in heavy-tailed settings. In tasks that involve constructing simultaneous confidence intervals for a large number of mean parameters, the proposed method offers more reliable level control than all the alternative methods for which comparable theoretical guarantees are available. In addition, we included comparisons with an approach based on self-normalization that we adapted from the paper (Liu and Shao, 2014), which dealt with a different set of inference problems. This adapted method provides somewhat tighter confidence intervals than the proposed method while maintaining similar level control, but is not supported by comparable theoretical guarantees in our context.

Looking ahead to future research, a natural and important question is to determine if near $n^{-1/2}$ rates of bootstrap approximation can be established for the method based on self-normalization. In addition, there are a number of other related questions dealing with model assumptions. For instance, it would be of interest to know if theoretical results for the proposed method or the self-normalization method can be established under weaker moment assumptions than those used here. Likewise, our analysis might be extended in other ways, such as by removing the presence of variance decay structure, or by allowing for samples to be corrupted according to a contamination model.

Supplementary Materials

The supplementary materials contain the proofs of Theorem 1 and Proposition 1.

Acknowledgements

We are grateful to Mengxin Yu for generously providing the code for the HL method.

References

Abdalla, P. and N. Zhivotovskiy (2022). Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *arXiv:2205.08494*.

Alpha Vantage (2024). Alpha Vantage API. https://www.alphavantage.co.

Bai, Z. and J. W. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Chang, J., Q.-M. Shao, and W.-X. Zhou (2016). Cramér-type moderate deviations for studentized two-sample U-statistics with applications.

Chen, X. and K. Kato (2020). Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields 176*, 1097–1163.

Chen, Y.-C., C. R. Genovese, and L. Wasserman (2015). Asymptotic theory for density ridges. *The Annals of Statistics 43*(5), 1896–1928.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*(6), 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics 42*(5), 1787–1818.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*(4), 2309–2352.

Chernozhukov, V., D. Chetverikov, and Y. Koike (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *The Annals of Applied Probability 33*(3), 2374–2425.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018). The econometrics of shape restrictions. *Annual Review of Economics 10*, 31–63.

Comon, P. and C. Jutten (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press.

de la Peña, V. H., T. L. Lai, and Q.-M. Shao (2009). *Self-normalized processes: Limit theory and Statistical Applications*. Springer.

Delaigle, A., P. Hall, and J. Jin (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology 73*(3), 283–301.

Deng, H. and C.-H. Zhang (2020). Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *The Annals of Statistics 48*(6), 3643–3671.

Dette, H., K. Kokot, and A. Aue (2020). Functional data analysis in the Banach space of continuous functions. *The Annals of Statistics 48*(2), 1168–1192.

Fan, J., P. Hall, and Q. Yao (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *Journal of the American Statistical Association 102*(480), 1282–1288.

Fan, J., Z. Lou, and M. Yu (2023). Robust high-dimensional tuning free multiple testing. *The Annals of Statistics 51*(5), 2093–2115.

Fang, X., Y. Koike, S.-H. Liu, and Y.-K. Zhao (2023). High-dimensional central limit theorems by Stein's method in the degenerate case. *arXiv:2305.17365*.

Foss, S., D. Korshunov, and S. Zachary (2011). *An Introduction to Heavy-tailed and Subexponential Distributions*. Springer.

Giessing, A. (2023). Gaussian and bootstrap approximations for suprema of empirical processes.

arXiv:2309.01307.

Giessing, A. and J. Fan (2020). Bootstrapping ℓ_p-statistics in high dimensions. *arXiv:2006.13099*.

Han, F., S. Xu, and W.-X. Zhou (2018). On Gaussian comparison inequality and its application to spectral analysis of large random matrices. *Bernoulli 24*(3), 1787–1833.

Hodges, J. and E. Lehmann (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics 34*(2), 598–611.

Johnstone, I. M. (2019+). *Gaussian Estimation: Sequence and Wavelet Models*. preprint.

Ke, Y., S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science 34*(3), 454–471.

Kock, A. B. and D. Preinerstorfer (2024). A remark on moment-dependent phase transitions in high-dimensional Gaussian approximations. *Statistics & Probability Letters 211*, 110149.

Kock, A. B. and D. Preinerstorfer (2025). High-dimensional Gaussian and bootstrap approximations for robust means. *arXiv:2504.08435*.

Koike, Y. (2024). High-dimensional bootstrap and asymptotic expansion. *arXiv:2404.05006*.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2019). *Continuous Multivariate Distributions, Volume 1: Models and Applications*, Volume 334. John Wiley & Sons.

Kuchibhotla, A. K., L. D. Brown, A. Buja, J. Cai, E. I. George, and L. H. Zhao (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics 48*(5), 2953–2981.

Kuchibhotla, A. K., S. Mukherjee, and D. Banerjee (2021). High-dimensional CLT: Improvements, non-uniform extensions and large deviations. *Bernoulli 27*(1), 192 – 217.

Kuchibhotla, A. K. and A. Rinaldo (2020). High-dimensional CLT for sums of non-degenerate random vectors: $n^{-1/2}$-rate. *arXiv:2009.13673*.

Liu, M. and M. E. Lopes (2024). Robust max statistics for high-dimensional inference. *arXiv:2409.16683*.

Liu, W. and Q.-M. Shao (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics 42*(5), 2003 – 2025.

Lopes, M. E. (2022). Central limit theorem and bootstrap approximation in high dimensions: Near $1/\sqrt{n}$ rates via implicit smoothing. *The Annals of Statistics 50*(5), 2492–2513.

Lopes, M. E., N. B. Erichson, and M. W. Mahoney (2023). Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching. *Bernoulli 29*(1), 428–450.

Lopes, M. E., Z. Lin, and H.-G. Müller (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *The Annals of Statistics 48*(2), 1214–1229.

Lou, Z. and W. B. Wu (2017). Simultaneous inference for high dimensional mean vectors. *arXiv:1704.04806*.

Lugosi, G. and S. Mendelson (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics 19*(5), 1145–1190.

Mendelson, S. and N. Zhivotovskiy (2020). Robust covariance estimation under L_4-L_2 norm equivalence. *The Annals of Statistics 48*(3), 1648–1664.

Muirhead, R. J. (2009). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons.

Nair, J., A. Wierman, and B. Zwart (2022). *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge.

Nemirovsky, A. S. and D. B. Yudin (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley.

Oksendal, B. (2013). *Stochastic Differential Equations: An Introduction with Applications*. Springer.

Pólya, G. (1949). Remarks on characteristic functions. In *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, pp. 115–123.

Resende, L. (2024). Robust high-dimensional Gaussian and bootstrap approximations for trimmed sample means. *arXiv:2410.22085*.

Roy, A., K. Balasubramanian, and M. A. Erdogdu (2021). On empirical risk minimization with dependent and heavy-tailed data. *Advances in Neural Information Processing Systems 34*, 8913–8926.

Ruppert, D. and D. S. Matteson (2011). *Statistics and Data Analysis for Financial Engineering with R Examples*. Springer.

Singh, R. and S. Vijaykumar (2023). Kernel ridge regression inference. *arXiv:2302.06578*.

Sun, Y., X. He, and J. Hu (2022). An omnibus test for detection of subgroup treatment effects via data partitioning. *Annals of Applied Statistics 16*(4), 2266–2278.

Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons.

Yu, M. and X. Chen (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(2), 247–270.

Zhang, D. and W. B. Wu (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics 45*(5), 1895–1919.

University of California, Davis

E-mail: (mshliu@ucdavis.edu)

University of California, Davis

E-mail: (melopes@ucdavis.edu)