Improving comparison of power and level for conservative multiplier bootstrap tests in simulation studies and beyond

Lyudmila Sakhanenko *
sakhanen@msu.edu

September 1, 2025

Abstract

We propose a simple modification procedure that helps to compare the levels and powers for conservative multiplier bootstrap tests. It is especially useful for simulation studies where empirical levels are zero. We provide a theoretical justification and illustrate the use of the procedure in a recent class of multiplier bootstrap tests for quantile regression and in a recent class of high-dimensional tests for MANOVA.

MSC2020: 62H15, 62F40, 62F99
Keywords: Multiplier Bootstrap, high-dimensional tests

1 Introduction

Bootstrap was introduced by Efron (1979). Since then, it has been widely applied in many settings where the explicit limit distribution is either unknown or intractable.

Multiplier bootstrap tests have received a renewed attention in the literature with the works of Chernozhukov et al. (2012, 2014, 2017) and several applications to testing in the high-dimensional setting when the dimension grows exponentially with the sample size; see for example Chen (2018), Chen and Zhou (2020), and Chakraborty and Sakhanenko (2023).

In practice on simulated and real datasets, these tests often suffer from being too conservative (their levels are below the nominal level) and their power tends to be on the lower side when the dimension is very large compared to the sample size. For example, as studied in Pan and Zhou (2021) the multiplier bootstrap tests for quantile regression based on both exponential and Rademacher weights had very low levels and rather mediocre power under dense alternative

*Department of Statistics and Probability, Michigan State University, 619 Red Cedar Rd, East Lansing, 48824, USA

across all considered distributions. As another motivating example, we consider the multiplier test for MANOVA, which was recently proposed and studied in Chakraborty and Sakhanenko (2023). This test was very conservative and had moderate power under some distributions such as t.

Statisticians customarily perform simulation studies in which each setup is repeated independently M times, while each m-th run consists of using a random sample $(X_1^{(m)}, \ldots, X_n^{(m)})$ from the setup to obtain a test statistic value $T_n^{(m)}$ and to construct a series of bootstrapped versions of the test statistic $T_{n,e}^{(m)}$ that also incorporate many independent samples (say 10K) of multipliers $(e_1^{(k,m)}, \ldots, e_n^{(k,m)}), k = 1, \ldots, 10K$, which in turn produce the quantiles $Q_n^{(m)}$ that are used for the test. Then the proportion of values $T_n^{(m)}$ that are, say larger, than $Q_n^{(m)}$ gives an empirical size or empirical power of the test according to the underlying hypothesis. When an empirical size is consistently smaller than the nominal size α a test is called conservative. Statisticians prefer tests that have empirical sizes close to the nominal and powers that are close to 1.

In this work we explore what happens if one would take all $Q_n^{(m)}, m = 1, \ldots, M$, quantiles and create a super-quantile, say Q_n^*, and then use it to estimate the empirical size and the empirical power of the test from the same datasets $(X_1^{(m)}, \ldots, X_n^{(m)}), m = 1, \ldots, M$. Since the data are becoming more abundant and it is becoming cheap to obtain repeated measurements in real-life experiments, this is still a widely encountered situation.

In Section 2 we show theoretically that using selected Q_n^* for conservative multiplier bootstrap tests improves the empirical size and empirical power. We also illustrate our theoretical findings with a simulation study in Section 3. We draw conclusions and raise new questions in Section 4.

2 Theoretical underpinnings

2.1 Test on an average

Let us start with a toy scenario. Let $(X_1, \ldots, X_n)$ be i.i.d. random variables with unknown mean μ and variance $\sigma^2 > 0$. For ease of presentation, consider a classical problem of testing for a specific mean $H_0 : \mu = \mu_0$ and consider a one-sided alternative $H_A : \mu > \mu_0$. The test statistic is the empirical mean $T_n = n^{-1} \sum_{i=1}^n X_i$. Consider a multiplier bootstrap for the limiting distribution of T_n. To this end, define $T_{n,e} = n^{-1} \sum_{i=1}^n e_i (X_i - T_n)$, where $(e_1, \ldots, e_n)$ is a random sample of multipliers from the standard normal distribution independent of $(X_1, \ldots, X_n)$. The conditional distribution of $T_{n,e}$ given $(X_1, \ldots, X_n)$ is normal with mean zero and variance $\frac{(n-1)}{n} \frac{S^2}{n}$, where $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - T_n)^2$. Introduce the $100(1-\alpha)$-th quantile $Q_n(\alpha)$ of the conditional distribution. Naturally, $Q_n(\alpha) = \sqrt{\frac{n-1}{n}} \frac{S}{\sqrt{n}} z_\alpha$, where $P(Z > z_\alpha) = \alpha$ for a standard normal variable Z. In practice, we use $\hat{Q}_n(\alpha)$, which is the empirical quantile of an empirical bootstrap distribution based on

the large number, say 10K, of repetitions of the multiplier samples.

Let $Z = \frac{T_n - \mu}{S/\sqrt{n}}$. Its distribution is asymptotically close to the Student t with $n-1$ degrees of freedom, and it has the standard normal limiting distribution under the usual moment assumptions. Therefore, the size of the test is

$$P(T_n > \mu_0 + Q_n(\alpha) | \mu = \mu_0) = P\bigg(Z > \sqrt{\frac{n-1}{n}} z_\alpha\bigg)$$

and the power is as follows for some $\Delta \neq 0$

$$P(T_n > \mu_0 + Q_n(\alpha) | \mu = \mu_0 + \Delta) = P\bigg(Z > -\frac{\sqrt{n}\Delta}{S} + \sqrt{\frac{n-1}{n}} z_\alpha\bigg).$$

Now consider M independent random samples $(X_1^{(m)}, \ldots, X_n^{(m)}), m = 1, \ldots, M$ from the same distribution. For each sample, obtain $T_n^{(m)}$ and $Q_n^{(m)}(\alpha)$. Consider empirical p-values defined as follows

$$\frac{1}{M} \sum_{k=1}^{M} I(T_n^{(k)} > Q_n^{(m)}(\alpha)) = p^{(m)}.$$

Perform the classical Benjamini and Hochberg's procedure (1995) to control the false discovery rate (FDR) at the level α. That is order the p-values, say $p_1 \leq \cdots \leq p_M$ and then find the largest i for which $p_i \leq \frac{i}{M}\alpha$. Since the empirical cumulative distribution function is non-decreasing, it is equivalent to finding the largest i for which

$$\sum_{k=1}^{M} I(T_n^{(k)} > \tilde{Q}_n^{(M-i+1)}) \leq i\alpha,$$

where $\tilde{Q}_n^{(1)} \leq \cdots \leq \tilde{Q}_n^{(M)}$ are the ordered quantiles. Then we define the super-quantile $Q_n^*(\alpha)$ as $\tilde{Q}_n^{(i)}$ for the best i from the FDR procedure. If the FDR procedure fails to find such index i then take $\tilde{Q}_n^{(1)}$, which is the smallest of all quantiles $Q_n^{(m)}(\alpha), m = 1, \ldots, M$.

Consider a test that rejects H_0 if $T_n^{(m)} > Q_n^*(\alpha)$. Then the size of this test is

$$P\bigg(Z > \sqrt{\frac{n-1}{n}} \frac{S^*}{S^{(m)}} z_\alpha | (X_1^{(m)}, \ldots, X_n^{(m)})\bigg),$$

while the power of this test is

$$P\bigg(Z > -\frac{\sqrt{n}\Delta}{S^{(m)}} + \sqrt{\frac{n-1}{n}} \frac{S^*}{S^{(m)}} z_\alpha | (X_1^{(m)}, \ldots, X_n^{(m)})\bigg).$$

Obviously, to improve the size one would like to choose $Q_n^*(\alpha)$ and therefore the corresponding S^* such that $\sqrt{\frac{n-1}{n}} \frac{S^*}{S^{(m)}}$ is close to 1. However, if one would

like to improve the power, one would like the smallest $\sqrt{\frac{n-1}{n}}\frac{S^*}{S^{(m)}}$, which would come from the smallest $Q_n^{(m)}$. This balance is achieved by the combination of FDR procedure and taking the smallest of quantiles $Q_n^{(m)}(\alpha)$.

On the event $\{S^* \neq S^{(m)}\}$ denote $b := S^{(m)} - S^*$, then the improved size is

$$\begin{aligned}\hat{\alpha}_n &= P\Big(Z > \sqrt{\frac{n-1}{n}}\frac{S^{(m)}-b}{S^{(m)}}z_\alpha | (X_1^{(m)},\dots,X_n^{(m)})\Big)\\ &= P(Z > \sqrt{\frac{n-1}{n}}z_\alpha)\\ &+ \frac{M-1}{M}P\Big(Z \in \Big[\sqrt{\frac{n-1}{n}}z_\alpha - \sqrt{\frac{n-1}{n}}\frac{b}{S^{(m)}}z_\alpha, \sqrt{\frac{n-1}{n}}z_\alpha\Big] | (X_1^{(m)},\dots,X_n^{(m)})\Big),\end{aligned}$$

where the first term is the size of the original test, say α_n. Similarly, the power of the new test, say $\hat{p}_n$ is equal to the power of the original test p_n plus the following

$$\frac{M-1}{M}P\Big(Z \in \Big[-\frac{\sqrt{n}\Delta}{S^{(m)}} - \sqrt{\frac{n-1}{n}}\frac{b}{S^{(m)}}z_\alpha, -\frac{\sqrt{n}\Delta}{S^{(m)}} + \sqrt{\frac{n-1}{n}}z_\alpha\Big] | (X_1^{(m)},\dots,X_n^{(m)})\Big)$$

Consider the asymptotical behavior of $\hat{\alpha}_n$ and $\hat{p}_n$ when $b \to 0^+$ in probability. Using calculus, we derive

$$\begin{aligned}\hat{\alpha}_n = &\quad \alpha_n + \frac{M-1}{M}\sqrt{\frac{n-1}{n}}\frac{b}{S^{(m)}}z_{\alpha/}f_{n-1}(\sqrt{\frac{n-1}{n}}z_\alpha)(1+o(1))\\ \hat{p}_n = &\quad p_n + \frac{M-1}{M}\sqrt{\frac{n-1}{n}}\frac{b}{S^{(m)}}z_\alpha f_{n-1}(-\frac{\sqrt{n}\Delta}{S^{(m)}} + \sqrt{\frac{n-1}{n}}z_\alpha),\end{aligned}$$

where

$$f_{n-1}(u) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)\sqrt{\pi(n-1)}}(1+u^2/(n-1))^{-n/2},\ u \in \mathbb{R}.$$

Using Taylor series for the function $(1+u^2/(n-1))^{-n/2}$ and Stirling's approximation $\Gamma(m) \approx \sqrt{2\pi}m^{m+0.5}e^{-m}$ we obtain as $n \to \infty$

$$\hat{\alpha}_n = \alpha + \frac{M-1}{M}\sqrt{\frac{1}{2\pi}}\frac{b}{\sigma}z_\alpha e^{-0.5z_\alpha^2}(1+o(1))$$

and

$$\hat{p}_n = p_n + \frac{M-1}{M}\sqrt{\frac{1}{2\pi}}\frac{b}{\sigma}z_\alpha^2\frac{\sqrt{n}\Delta}{S^{(m)}}(1+o(1)).$$

Note that the power improvement for a conservative test would be

$$\frac{\sqrt{n}\Delta}{S^{(m)}}z_\alpha e^{0.5z_\alpha^2}(\hat{\alpha}_n - \alpha_n),$$

where Δ is getting close to 0 for local alternatives. The gain in power could be useful and substantive.

2.2 General setting

Now consider a more general setting with a generic null hypothesis H_0. Let $(X_1, \ldots, X_n)$ be i.i.d. random vectors in $\mathbb{R}^d$. Let a test statistic be defined as

$$T_n = \max_{\theta \in \Theta} \sum_{i=1}^{n} \Phi(X_i; \theta)$$

and let its bootstrapped version be defined as

$$T_n^e = \max_{\theta \in \Theta} \sum_{i=1}^{n} e_i \Phi(X_i; \theta),$$

where $\theta \in \Theta$ represents a set of tuning parameters such as the index $j = 1, \ldots, d$; $e_1, \ldots, e_n$ are i.i.d. multipliers independent of $(X_1, \ldots, X_n)$ from a known distribution with mean 0 and variance 1, such as standard normal. Then a multiplier bootstrap test rejects H_0 at the significance level $\alpha \in (0, 1)$ if

$$T_n \geq Q_n(\alpha) := \inf\{u \in \mathbb{R} : P_e(T_n^e \leq u) \geq 1 - \alpha\}.$$

These tests are especially useful for symmetry testing and various hypothesis about the shape of the distribution when a parametric bootstrap does not work. See Kosorok (2008).

Now consider M independent random samples $(X_1^{(m)}, \ldots, X_n^{(m)}), m = 1, \ldots, M$ from the same distribution. For each sample, obtain $T_n^{(m)}$ and $Q_n^{(m)}(\alpha)$. Define the super-quantile $Q_n^*(\alpha)$ according to the following procedure.

1. Compute empirical p-values defined as follows

$$\frac{1}{M} \sum_{k=1}^{M} I(T_n^{(k)} > Q_n^{(m)}) = p^{(m)}.$$

2. Order the p-values, say $p_1 \leq \cdots \leq p_M$.

3. Find the largest i for which $p_i \leq \frac{i}{M}\alpha$. If such $1 \leq i \leq M$ exists, take the corresponding quantile as $Q_n^*(\alpha)$, else $Q_n^*(\alpha) := \min_{m=1,\ldots,M} Q_n^{(m)}(\alpha)$.

Now consider a test that rejects H_0 if $T_n \geq Q_n^*(\alpha)$. By the construction as $n \to \infty$ and $M \to \infty$ we have

$$P(T_n \geq Q_n^*(\alpha)) \geq P(T_n \geq Q_n^{(m)}(\alpha)).$$

Asymptotically, the FDR procedure would fail to find index i and the super-quantile would be the smallest of the quantiles $Q_n^{(m)}(\alpha)$. For a moderately large M one might wonder how would the FDR procedure play out. Note that the event $p_1 \leq \alpha$ is equivalent to asking for a binomial random variable B with parameters (M, α) to satisfy $B \leq \alpha$, which is equivalent to event $\{B = 0\}$.

Thus, this would happen in simulations with probability $(1-\alpha)^M$, which is rather small. Similarly, this situation would continue for $p_2, p_3, \ldots, p_{[1/\alpha]}$, where $[u]$ stands for the largest integer that is smaller than u. Then the next event $p_{[1/\alpha]+1} \leq ([1/\alpha]+1)\alpha$ is equivalent to $\{B = 0, 1\}$, which would happen in simulations with probability $(1-\alpha)^M + M(1-\alpha)^{M-1}\alpha$, which is still rather small. And so on, until we reach the last event $p_M \leq M\alpha$, which is equivalent to $\{B \leq M\alpha\}$. Using a normal approximation to the binomial distribution, this event has a probability of about 0.5. However, these binomial calculations are done separately, so we ignore that the conditions in step 3 of the FDR procedure are nested. So, it is quite unlikely that the FDR procedure here would give something different than $\min_{m=1,\ldots,M} Q_n^{(m)}(\alpha)$ unless the original test is conservative and actually attains the level $\alpha_0 << \alpha$. Then in all those probabilities we would replace α with α_0 while still checking step 3 with $p_i \leq i\alpha$. Then for medium size M and very small α_0 these probabilities are not close to 0 and FDR procedure would yield a quantile different from $\min_{m=1,\ldots,M} Q_n^{(m)}(\alpha)$, which is what we will observe in the simulation study next.

3 Simulation study

To illustrate the proposed improvement for multiplier bootstrap tests, we consider two examples that recently appeared in the literature.

3.1 Multiplier bootstrap for a quantile regression

Consider the problem of testing the null hypothesis $H_0 : \beta_j^* = 0, j = 1, \ldots, d$ for the regression model $Y_i = \langle x_i, \beta^* \rangle + \sigma(x_i)\varepsilon_i, i = 1, \ldots, n$, where $x_i \in \mathbb{R}^d, i = 1, \ldots, n$, and $\varepsilon_i \in \mathbb{R}, i = 1, \ldots, n$, are independent. The later random variables are centered.

Pan and Zhou (2021) introduced and studied multiplier bootstrap tests based on Rademacher weights and exponential weights. They used the loss function

$$L_n(\beta) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle x_i, \beta^* \rangle)[0.5 - I(Y_i - \langle x_i, \beta^* \rangle < 0)]$$

to construct the test statistic as

$$T_n = L_n(0) - \min_{\beta \in \mathbb{R}^d} L_n(\beta).$$

For i.i.d. random weights $w_i, i = 1, \ldots, n$, the bootstrapped loss function

$$L_n^b(\beta) := \frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \langle x_i, \beta^* \rangle)[0.5 - I(Y_i - \langle x_i, \beta^* \rangle < 0)]$$

was then used to construct the bootstrapped test statistic

$$T_n^b = L_n^b(\hat{\beta}) - L_n^b(\hat{\beta}^b), \hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^d} L_n(\beta), \hat{\beta}^b = \text{argmin}_{\beta \in \mathbb{R}^d} L_n^b(\beta).$$

Then their test rejected H_0 at the significance level $\alpha \in (0,1)$ if

$$T_n > \inf\{r \in \mathbb{R} : P(T_n^b > r | (Y_i, x_i), i = 1, \ldots, n) \leq \alpha\}.$$

We took their setup for the simulation study and considered a homoscedastic model

$$Y_i = \langle x_i, \beta^* \rangle + \varepsilon_i, i = 1, \ldots, n,$$

and a heteroscedastic model

$$Y_i = \langle x_i, \beta^* \rangle + \frac{2 \exp x_{i1}}{1 + \exp\{x_{i1}\}} \varepsilon_i, i = 1, \ldots, n,$$

with errors ε_i coming from the t distribution with 2 degrees of freedom or one of 2 normal mixtures. The normal mixture of type I was defined as $\varepsilon_i = az_1 + (1-a)z_2$, where a is a Bernoulli random variable with probability 0.5 independent of 2 independent random normal variables $z_1 \sim N(-1, 1)$ and $z_2 \sim N(1, 1)$. The normal mixture of type II was based on a Bernoulli random variable a with probability 0.9 independent of 2 independent random normal variables $z_1 \sim N(0, 1)$ and $z_2 \sim N(0, 5^2)$.

Pan and Zhou (2021) also used 3 different random designs for $x_i \sim N(0, \Sigma)$, which we also consider. The independent design used $\Sigma = \mathbb{I}_d$, while weakly correlated design generated the off-diagonal components of the covariance matrix $\sigma_{jk} = 0.5^{|j-k|}(\sigma_{jj}\sigma_{kk})^{1/2}$ from the diagonal components $\sigma_{jj} \sim U(0.5, 1)$, which were distributed independently. The equally correlated design generated off diagonal components as $\sigma_{jk} = 0.5(\sigma_{jj}\sigma_{kk})^{1/2}$ from the same sample of diagonal components that were independent and uniformly distributed on [0.5, 1].

Just as Pan and Zhou (2021) we use two sets of bootstrap weights: exponential ones produce the test mb-exp and Rademacher ones produce the test mb-Rad. The names were introduced in Pan and Zhou (2021).

We used $n = 200, d = 15$, 1K bootstrap samples and repeated each scenario $M = 300$ times to obtain empirical levels and empirical powers for the mb-exp and mb-Rad tests in Pan and Zhou (2021). Two alternatives were used: sparse H_A with $\beta_1^* = 0.5$ while the rest of βs are zero and dense H_A with $\beta_j^* = 0.1, j = 1, \ldots, 10$ while the rest of βs are zero.

We then applied our procedure to those M bootstrapped samples of the test statistics and the bootstrapped quantiles $T_n^{(m)}, Q_n^{(m)}, m = 1, \ldots, M$, to obtain improved levels and powers. The results are reported in Table 1.

The levels and the powers are consistently above those obtained via multiplier tests by Pan and Zhou (2021). In some cases the power improvement is as dramatic as going from .21 to .813 (homoscedastic model with errors from normal mixture of type II under independent normal design for sparse alternative). In other cases the improvement makes the test worthwhile as in the case of dense alternative under a model with errors from a normal mixture of type II.

Table 1: Comparative performance for the original test in Pan and Zhou and its super-quantile version, the numbers for the later are given in brackets. The nominal test level is 0.01.

model	H_0 mb-exp	H_0 mb-Rad	sparse H_A mb-exp	sparse H_A mb-Rad	dense H_A mb-exp	dense H_A mb-Rad
t homo, ind N	.000(.007)	.005(.007)	.535(.937)	.600(.967)	.075(.463)	.085(.513)
t homo, weakly cor N	.000(.007)	.005(.007)	.365(.710)	.370(.773)	.375(.910)	.405(.727)
t homo, equally cor N	.000(.007)	.005(.007)	.345(.817)	.380(.800)	.850(1.00)	.845(1.00)
t hetero, ind N	.000(.007)	.000(.007)	.770(.993)	.810(.997)	.160(.553)	.200(.480)
t hetero, weakly cor N	.000(.007)	.000(.007)	.490(.910)	.510(.907)	.475(.803)	.535(.770)
t hetero, equally cor N	.000(.007)	.000(.007)	.450(.923)	.500(.947)	.920(1.00)	.930(1.00)
N mix I, homo, ind N	.000(.007)	.005(.007)	.295(.700)	.330(.700)	.045(.220)	.075(.230)
N mix I, homo, weakly cor N	.005(.007)	.007(.007)	.230(.453)	.250(.410)	.210(.473)	.230(.490)
N mix I, homo, equally cor N	.007(.007)	.007(.007)	.205(.933)	.225(.940)	.595(.960)	.615(.967)
N mix I, hetero, ind N	.000(.007)	.000(.007)	.475(.980)	.510(.990)	.095(.190)	.110(.233)
N mix I, hetero, weakly cor N	.005(.007)	.005(.007)	.330(.743)	.365(.737)	.260(.747)	.295(.613)
N mix I, hetero, equally cor N	.005(.007)	.005(.007)	.270(.840)	.300(.887)	.715(.987)	.735(.987)
N mix II, homo, ind N	.003(.007)	.003(.007)	.210(.813)	.170(.773)	.010(.193)	.010(.193)
N mix II, homo, weakly cor N	.005(.007)	.005(.007)	.163(.573)	.150(.707)	.093(.510)	.100(.503)
N mix II, homo, equally cor N	.005(.007)	.005(.007)	.153(.867)	.123(.897)	.527(.940)	.487(.953)
N mix II, hetero, ind N	.005(.007)	.005(.007)	.190(.573)	.177(.563)	.033(.217)	.027(.237)
N mix II, hetero, weakly cor N	.007(.007)	.007(.007)	.123(.663)	.107(.733)	.077(.500)	.083(.433)
N mix II, hetero, equally cor N	.000(.007)	.000(.007)	.113(.733)	.100(.573)	.437(.963)	.433(.953)

3.2 Multiplier bootstrap tests for high-dimensional data for MANOVA

Consider the classical problem of testing the hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \text{ vs } H_A : \text{ otherwise,}$$

where there are K independent groups of random vectors $V_{k,i} \in \mathbb{R}^p, i = 1, \ldots, n, k = 1, \ldots, K$, drawn from K populations with unknown means $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$.

For the i-th vector in the k-th group $V_{k,i}$, its components are denoted by $[V_{k,i}]_q, q = 1, \ldots, p$. For ultra-high dimension p Chakraborty and Sakhanenko (2023) proposed a multiplier bootstrap approach. Their test statistic was given by

$$T_n = \max_{l=1,\ldots,L} \max_{j=1,\ldots,d} n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{q=1}^{p} A^{(l)}_{j(q+(k-1)p)} [V_{k,i}]_q,$$

where the matrices $A^{(l)}, l = 1, \ldots, L$, satisfied some sparsity and null hypothesis conditions.

Chakraborty and Sakhanenko (2023) then proposed the bootstrapped test statistics as follows

$$T^e_n = \max_{l=1,\ldots,L} \max_{j=1,\ldots,d} n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{q=1}^{p} A^{(l)}_{j(q+(k-1)p)} e_i [V_{k,i} - \bar{V_k}]_q,$$

where the vector $(e_1, \ldots, e_n)$ of iid $N(0, 1)$ random variables is independent of all $V_{k,i}$. Then their test rejected H_0 at the significance level $\alpha \in (0, 1)$ if

$$T_n > Q_\alpha := \inf\{u \in \mathbb{R} : P(T^e_n \leq u | V_{k,i}, i = 1, \ldots, n, k = 1, \ldots, K) \geq 1 - \alpha.$$

We consider the same setup as what they used in the simulation study. $K = 4$ and $V_{t,i} = \mu_t + \Gamma Z_{t,i}, t = 1, 2, 3, 4; i = 1, \ldots, n_t$, where $Z_{t,i}$ were generated from one of 3 models. The case of underlying t_4 distribution of $V_{k,i}$ was quite challenging in their simulation study. Let us apply the proposed improvement scheme to this case and to their test example reported in their Table 1 for their test $T^{(2)}_n$, which is based on 5-diagonal matrices $A^{(l)}$.

The sample sizes were $\mathbf{n_1} = (25, 30, 40, 50)$, $\mathbf{n_2} = (50, 60, 80, 100)$, and $\mathbf{n_3} = (100, 120, 160, 200)$, while for the the dimension of the data we picked $p = 500$ and $p = 1000$. They considered covariance matrix $(1 - \rho)\mathbb{I}_p + \rho\mathbb{J}_p$, where $\mathbb{I}_p$ stands for the identity $p \times p$ matrix and $\mathbb{J}_p$ denotes $p \times p$ matrix of ones. The parameter ρ took values 0.1, 0.5, 0.9. They also considered the covariance of the form $0.6^{|i-j|}, i, j, = 1, \ldots, p$, denoted by ρ =NA. We report the levels and powers in Table 2.

Based on the Table 2 the super-quantile test has levels that are closer to the nominal and much better powers than the original test. The best gains in level happen where the original test had levels close to 0 such as $\rho = NA$ and $\rho = 0.1$ cases, corresponding to a non-linear covariance structure and almost a spherical covariance structure, respectively. The gains in level and power are somewhat moderate for the case of $\rho = 0.9$ when the covariance matrix is close to a singular matrix $\mathbb{J}_p$.

Table 2: Comparative performance for original test and super-quantile test, the numbers for the later are given in brackets. The nominal test level is 0.05.

p	ρ	$\mathbf{n_1}$ size	$\mathbf{n_1}$ power	$\mathbf{n_2}$ size	$\mathbf{n_2}$ power	$\mathbf{n_3}$ size	$\mathbf{n_3}$ power
500	.1	.007(.049)	.122(.413)	.016(.046)	.271(.507)	.025(.048)	.254(.402)
	.5	.027(.047)	.122(.308)	.029(.053)	.191(.338)	.029(.050)	.164(.251)
	.9	.037(.062)	.163(.338)	.054(.054)	.005(.335)	.043(.054)	.161(.233)
	NA	.001(.047)	.005(.059)	.012(.028)	.018(.403)	.002(.026)	.190(.315)
1000	.1	.004(.031)	.082(.342)	.006(.048)	.165(.384)	.024(.054)	.290(.488)
	.5	.020(.051)	.097(.252)	.024(.048)	.113(.255)	.034(.051)	.155(.268)
	.9	.041(.073)	.155(.307)	.042(.065)	.146(.258)	.042(.056)	.154(.233)
	NA	.003(.009)	.031(.220)	.008(.045)	.088(.278)	.010(.047)	.165(.324)

3.3 Comparison of the (n, M) study with the (n', M') study while $nM = n'M'$

In this section we would check using Example 1 setup what happens if we split the data differently. Indeed, overall we use nM data points for the study when we compute a super-quantile. What if we would use less repetitions $M' < M$ and larger samples $n' > n$ while keeping $nM = n'M'$? Maybe the original test would do great with n' and would beat the proposed test based on a super-quantile. This question is at the heart of the most obvious criticism of the proposed approach.

To investigate this question we consider $nM = 300n$ split onto $M' = 100$ and $n' = 3n$. The results are summarized in Table 3. The levels are too close to call, since $M' = 100$ with $\alpha = 0.01$. The powers of the super-quantile test are significantly higher for the dense alternative for all the cases except equally correlated design where they are almost the same. The powers of the super-quantile test are significantly higher for sparse alternative when errors come from a balanced normal mixture (type I) and the design is independent or equally correlated (with a homogeneous model). When errors are from unbalanced normal mixture (type II) or t distribution, the powers of the super-quantile test are significantly higher under sparse alternative for all the scenarios except independent design for heterogeneous model. In all the other cases (13 out of 36), both tests have similar powers.

Using the same amount of total data (nM) the super-quantile test is a valuable option to employ and would get to a correct rejection more often than a typical bootstrap multiplier test. So in real applications, one might consider splitting the dataset with $n, M = 1$ into a few $M' > 1$ smaller samples with $n'M' = n$ and running super-quantile test. This could be quite attractive for large n datasets. Finally, there is no cost of generating multiplier random samples compared to costs collecting real data.

Table 3: Comparative performance for the original test in Pan and Zhou with $n = 600, M = 100$ against our test with $n = 200, M = 300$, the numbers for the later are given in brackets. The nominal test level is 0.01.

model	H_0 mb-exp	H_0 mb-Rad	sparse H_A mb-exp	sparse H_A mb-Rad	dense H_A mb-exp	dense H_A mb-Rad
t homo, ind N	.00(.007)	.00(.007)	.82(.937)	.85(.967)	.08(.463)	.06(.513)
t homo, weakly cor N	.02(.007)	.01(.007)	.75(.710)	.74(.773)	.61(.910)	.61(.727)
t homo, equally cor N	.01(.007)	.01(.007)	.76(.817)	.74(.800)	1.00(1.00)	1.00(1.00)
t hetero, ind N	.00(.007)	.00(.007)	.90(.993)	.87(.997)	.160(.553)	.18(.480)
t hetero, weakly cor N	.00(.007)	.00(.007)	.90(.910)	.90 (.907)	.49(.803)	.49(.770)
t hetero, equally cor N	.01(.007)	.00(.007)	.90(.923)	.90(.947)	1.00(1.00)	1.00(1.00)
N mix I, homo, ind N	.02(.007)	.02(.007)	.62(.700)	.63(.700)	.04(.220)	.03(.230)
N mix I, homo, weakly cor N	.00(.007)	.00(.007)	.40(.453)	.39(.410)	.21(.473)	.24(.490)
N mix I, homo, equally cor N	.00(.007)	.01(.007)	.57(.933)	.60(.940)	.96(.960)	.96(.967)
N mix I, hetero, ind N	.01(.007)	.01(.007)	.75(.980)	.77(.990)	.10(.190)	.10(.233)
N mix I, hetero, weakly cor N	.00(.007)	.00(.007)	.78(.743)	.75(.737)	.54(.747)	.53(.613)
N mix I, hetero, equally cor N	.00(.007)	.00(.007)	.82(.840)	.82(.887)	.98(.987)	.98(.987)
N mix II, homo, ind N	.00(.007)	.00(.007)	.61(.813)	.63(.773)	.09(.193)	.10(.193)
N mix II, homo, weakly cor N	.01(.007)	.01(.007)	.42(.573)	.43(.707)	.31(.510)	.32(.503)
N mix II, homo, equally cor N	.00(.007)	.00(.007)	.59(.867)	.53(.897)	.95(.940)	.94(.953)
N mix II, hetero, ind N	.00(.007)	.00(.007)	.57(.573)	.56(.563)	.09(.217)	.09(.237)
N mix II, hetero, weakly cor N	.00(.007)	.01(.007)	.50(.663)	.52(.733)	.34(.500)	.22(.433)
N mix II, hetero, equally cor N	.01(.007)	.00(.007)	.53(.733)	.59(.573)	.98(.963)	.98(.953)

4 Summary and discussion

We propose a procedure that modifies levels and powers of conservative multiplier bootstrap tests. The procedure harnesses bootstrap quantiles from several independent random samples from the same setting to create a super- quantile by using FDR control and minimization. It works naturally in simulation studies. There this procedure can be employed for a more meaningful comparison of conservative tests. One can compare improved levels against improved levels and improved powers against improved powers as opposed to looking at zero empirical levels for both tests.

In one extreme, it would require M rather than 1 random sample for any real application, which is not a severe constraint given the abundance of data in the current era of data science. Alternatively, one can split the original data but generate more random samples of multipliers to perform the modification and obtain meaningful empirical levels for conservative tests.

References

[1] Benjamini, Y., Hochberg, Y. (1995) Controlling the False Discovery rate: a practical and powerful approach to multiple testing, J.R. Statist. Soc. B, **57** (1), pp. 289–300.

[2] Chakraborty, N., Sakhanenko, L. (2023) Novel multiplier bootstrap tests for high-dimensional data with applications to MANOVA, *Computational Statistics and Data Analysis*, Vol 178, https://doi.org/10.1016/j.csda.2022.107619

[3] Chen, X. (2018) Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Ann. Statist.* **46(2)**, pp. 642–678.

[4] Chen, X., Zhou, W.-X. (2020) Robust inference via multiplier bootstrap. *Ann. Statist.* **48(3)**, pp. 1665–1691.

[5] Chernozhukov, V., Chetverikov, D., Kato, K. (2017) Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45(4)**, pp. 2309–2352.

[6] Chernozhukov, V., Chetverikov, D., Kato, K. (2014) Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **46**, pp. 1564–1597.

[7] Chernozhukov, V., Chetverikov, D. Kato, K. (2012) Central limit theorems and multiplier bootstrap when p is much larger than n, *cemmap working paper*, **No. CWP45/12**, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.4512

[8] Efron, B. (1979) Bootstrap Methods: Another look at Jackknife, *Ann. Statist.*, **7(1)**, pp. 1–26.

[9] Kosorok, M. (2008) *Introduction to empirical processes and semiparametric inference*, Springer, New York.

[10] Pan, X., Zhou, W.-X. (2021) Multiplier bootstrap for quantile regression: non-asymptotic theory under random design. A Journal of the IMA, **10**, pp. 813–861.