CLUSTER-ROBUST INFERENCE ROBUST TO LARGE NON-IGNORABLE CLUSTERS

HAROLD D. CHIANG, YUYA SASAKI, AND YULONG WANG

Abstract. The recent literature points out that the conventional cluster-robust standard errors fail in the presence of large clusters. We propose a novel cluster-robust score subsampling inference method that is valid even in the presence of large clusters. Specifically, we derive the asymptotic distribution for the t-statistics based on the common cluster-robust variance estimators when the distribution of cluster sizes follows a power law with an exponent less than two. We then propose an inference procedure based on score subsampling and show its validity. Additionally, we prove that the wild cluster bootstrap is inconsistent under this environment. Our proposed method does not require tail index estimation and remains valid under the usual thin-tailed scenarios as well.

Keywords: cluster-robust inference, heavy-tailed distributions, large clusters, non-standard asymptotics, subsampling

Date: First arXiv version: August 19th, 2023.

We thank Bruce Hansen and Jack Porter for their constructive comments. H. Chiang's research is supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation. All remaining errors are ours.

1. Introduction

Consider the linear model

$$Y_{gi} = X_{gi}'\theta + U_{gi}, \quad \mathbb{E}[U_g|X_g] = 0,$$

where $X_g = (X_{g1}, \ldots, X_{gN_g})'$, $U_g = (U_{g1}, \ldots, U_{gN_g})'$, $g \in \{1, \ldots, G\}$ index clusters, and N_g denotes the size of the g-th cluster. Define the OLS estimator and the cluster-robust (CR) variance estimator by

$$\begin{aligned} \widehat{\theta} &= \left(\sum_{g=1}^{G}\sum_{i=1}^{N_g} X_{gi}X_{gi}'\right)^{-1} \sum_{g=1}^{G}\sum_{i=1}^{N_g} X_{gi}Y_{gi} = \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} \sum_{g=1}^{G}(X_g'X_g\theta + S_g) \quad \text{and} \\ \widehat{V}^{\mathrm{CR}} &= a_G \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} \left(\sum_{g=1}^{G} \widehat{S}_g\widehat{S}_g'\right) \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1}, \end{aligned}$$

respectively, for some finite sample adjustment $a_G \to 1$ as the number of clusters $G \to \infty$, where $S_g = \sum_{i=1}^{N_g} X_{gi}U_{gi}$, $\widehat{S}_g = \sum_{i=1}^{N_g} X_{gi}\widehat{U}_{gi}$, and $\widehat{U}_{gi} = Y_{gi} - X_{gi}'\widehat{\theta}$. For simplicity, we set $a_G = 1$ throughout.

Consider a linear combination, $\delta = r'\theta$, $r \in \mathbb{R}^{\dim(\theta)}$ with $\|r\| = 1$, as the parameter of interest. Let the corresponding estimator and the cluster-robust standard error be denoted by

$$\begin{aligned} \widehat{\delta} &= r'\widehat{\theta} \quad \text{and} \\ \widehat{\sigma}^2 &= r' \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} \left(\sum_{g=1}^{G} \widehat{S}_g\widehat{S}_g'\right) \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} r, \end{aligned}$$

respectively. A researcher is interested in conducting inference for δ using the common t-statistic

$$\frac{(\widehat{\delta} - \delta)}{\widehat{\sigma}} = \frac{r'(\widehat{\theta} - \theta)}{\sqrt{r' \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} \left(\sum_{g=1}^{G} \widehat{S}_g\widehat{S}_g'\right) \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} r}}$$

with the CR standard error.

Suppose each cluster size N_g is a random variable such that its distribution follows a power law with an exponent that is potentially less than two, and so the existence of $\mathbb{E}[\|S_g\|^2]$ is not guaranteed. The recent literature points out that the conventional cluster-robust inference fails in the presence of large clusters. It has been shown in Sasaki and Wang (2022) that if the cluster size distribution follows a power law with a less than two exponent, normal approximation-based inference can fail. In a different but closely related setup, Kojevnikov and Song (2023) show consistent estimation for variance estimation fails when there is only one large cluster. As an alternative, this paper proposes a novel cluster-robust score subsampling inference procedure that is also robust to heavy-tailed distributed cluster sizes. Estimation of the tail index for cluster size distribution is not required for the proposed inference procedure.

1.1. **Relations to the Literature.** Cluster-robust inference has a long history. Instead of attempting to exhaustively list the large body of the literature, we refer the reader to the surveys by, for example, Cameron and Miller (2015) and MacKinnon, Nielsen, and Webb (2023) for comprehensive reviews. The sampling frameworks in which cluster sizes themselves are random have been previously investigated by Bugni, Canay, Shaikh, and Tabord-Meehan (2022) and Sasaki and Wang (2022). Our key distributional approximation results for self-normalized sums are due to Logan, Mallows, Rice, and Shepp (1973), LePage, Woodroofe, and Zinn (1981), and Giné, Götze, and Mason (1997). For theoretical details of the underlying foundations of probability and statistics for heavy-tailed distributions, we refer the reader to Resnick (1987) and Resnick (2007). For the failure of empirical bootstrap for means of heavy-tailed distributions, see, e.g. Athreya (1987); Arcones and Giné (1989); Knight (1989). Our inference procedure relies on the theory of resampling method developed in Politis and Romano (1994) and Romano and Wolf (1999). Also, see Politis, Romano, and Wolf (1999) for a comprehensive treatment.

2. Robust Inference

We propose a novel cluster-robust score subsampling procedure that is also robust to heavy-tailed distributions of cluster sizes N_g. Our objective is to conduct statistical inference for δ using the self-normalized t-statistic $(\widehat{\delta} - \delta)/\widehat{\sigma}$. Denote the CDF of the sampling distribution of the t-statistic by

$$J_G^*(t) = \mathbb{P}\left((\widehat{\delta} - \delta)/\widehat{\sigma} \leqslant t\right).$$

It will be shown that, under appropriate conditions, it converges to the CDF of a limiting distribution $J^*(t)$. Consider a sequence of subsample sizes $b = b_G$ that grows with $b/G = o(1)$ as $G \to \infty$. Let $B_G = \binom{G}{b}$ denote the total possible number of subsamples of b clusters. For a given b and $j \in \{1, ..., B_G\}$, let $S_j \subset \{1, .., G\}$ be one of the B_G subsamples of the cluster indices with $|S_j| = b$, and define the score-resampled estimators

$$\begin{aligned} \widehat{\delta}_{b,j} =& r'\widehat{\theta}_{b,j} = \left(\frac{G}{b}\right) r' \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \sum_{g \in S_j} X_g' Y_g \qquad \text{and} \\ \widehat{\sigma}_{b,j}^2 =& \left(\frac{G}{b}\right)^2 r' \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\sum_{g \in S_j} \widehat{S}_{g,j} \widehat{S}_{g,j}'\right) \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} r, \end{aligned}$$

where $\widehat{S}_{g,j} = X_g'(Y_g - X_g \widehat{\theta}_{b,j})$. Note, in these definitions, that the inverse factor $(\sum_{g=1}^{G} X_g' X_g)^{-1}$ is calculated based on the full sample while the linear component and its variance are computed based on the subsample S_j – see Remark 2 below.

Define the empirical CDF of $(\widehat{\delta}_{b,j} - \widehat{\delta})/\widehat{\sigma}_{b,j}$ based on all possible B_G-subsamples by

$$L_{G,b}^*(t) = \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1}\left((\widehat{\delta}_{b,j} - \widehat{\delta})/\widehat{\sigma}_{b,j} \leqslant t\right).$$

It will be shown that, under suitable conditions, one can approximate $J^*(\cdot)$ by $L_{G,b}^*(\cdot)$ uniformly as the number G of clusters grows. In practice, however, $L_{G,b}^*(t)$ is computationally infeasible when G and b are both large. Thus, we randomly draw M such subsamples of clusters of size b with replacement, and define

$$\widehat{L}_{G,b}(t) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}\left((\widehat{\delta}_{b,j} - \widehat{\delta})/\widehat{\sigma}_{b,j} \leqslant t\right).$$

As M grows with the number G of clusters, this $\widehat{L}_{G,b}(\cdot)$ can be used in place of $L^{*}_{G,b}(\cdot)$.

For any $a \in (0,1)$, define the critical value

$$\widehat{c}_{G,b}(1-a) = \inf\left\{t \in \mathbb{R} : \widehat{L}_{G,b}(t) \geqslant 1-a\right\}.$$

Since $J^{*}(\cdot)$ has no point mass as we shall show, it follows that

$$\mathbb{P}\left((\widehat{\delta} - \delta)/\widehat{\sigma} \leqslant \widehat{c}_{G,b}(1-a)\right) \to 1-a$$

as $G \to \infty$. Therefore, this critical value leads to theoretically valid tests. In addition, a confidence region can be obtained by test-inversion.

Practical Implication: For the estimator $\widehat{\delta}$, one can continue to use the conventional cluster-robust "standard error" $\widehat{\sigma}$.[1] However, instead of using the conventional cutoff values, $\Phi^{-1}(0.025) \approx -1.96$ and $\Phi^{-1}(0.975) \approx 1.96$, one should use $\widehat{c}_{G,b}(0.025)$ and $\widehat{c}_{G,b}(0.975)$ obtained by subsampling to construct a 95% confidence interval for example.

Remark 1 (Tail index estimation is not required)**.** Unlike many other situations that involve statistical analysis of heavy-tailed distributions, our inference procedure does not require the estimation of the unknown tail index, which is a practical advantage.

Remark 2 (Finite sample non-invertibility of other cluster-based resampling methods)**.** In finite samples, when regressors contain a cluster-specific binary treatment variable or other dummies variables that are highly correlated within a cluster, $\sum_{g \in S_j} X_g' X_g$ might be singular especially for a small $b = |S_j|$, and thus the subsampled OLS might not be well-defined

[1] Note that the "standard error" $\widehat{\sigma}$ is not guaranteed to be consistent in the presence of large clusters. It may even diverge.

for a non-negligible proportion of subsamples. This issue is also faced by other cluster-based resampling methods, such as jackknife and bootstrap. In practice, several *ad hoc* "fixes," such as the use of generalized inverse or dropping such realizations, are employed, but their theoretical implications remain largely unclear. Our cluster-robust score subsampling procedure avoids such an issue in a theoretically supported manner.

2.1. **The Main Theoretical Result.** We start by briefly reviewing some definitions and known facts about stable distributions. For more detail, see Feller (1971), Zolotarev (1986), Samorodnitsky and Taqqu (1994), Embrechts, Klüppelberg, and Mikosch (1997), and Geluk and de Haan (2000) for example. A random variable η is said to be stable if it has a domain of attraction in that there exists a sequence of i.i.d. random variables $\xi_1, \xi_2, \ldots$ and sequences of positive numbers A_G and real numbers D_G such that as $G \to \infty$,

$$\frac{\sum_{g=1}^{G} \xi_g - D_G}{A_G} \xrightarrow{d} \eta.$$

A function $L(\cdot)$ is said to be slowly varying at ∞ if $\lim_{t\to\infty} L(yt)/L(t) = 1$ for all $y > 0$. If η is stable, then A_G takes the form of $G^{1/\alpha}L(G)$ for some $\alpha \in (0, 2]$ and some slowly varying function $L(\cdot)$ at ∞. In addition, if $\alpha \in (1, 2]$, then D_G can be chosen to be $G \cdot \mathbb{E}[\xi_g]$. Otherwise, one can set $D_G = 0$. Here, the number α is called the index of stability, and η is said to be α-stable. In such a case, ξ_g is said to belong to the domain of attraction of an α-stable distribution. We shall focus on the case with $\alpha \in (1, 2]$, since even the first moment of ξ_g would not be well-defined otherwise.

Assumption 1. $(X_g', S_g)_{g=1}^{G}$ are i.i.d., $\mathbb{E}[N_g] = c \in (0, \infty)$, and the design matrix satisfies

$$\left(\frac{1}{G} \sum_{g=1}^{G} X_g' X_g \right)^{-1} = Q^{-1} + o_p(1)$$

for a finite and positive definite matrix Q. In addition, for $v = r'Q^{-1}$ and for all $u_1, u_2 \in \mathbb{R}^{\dim(\theta)}$ with unit length, $v'S_g$ and $u_1'X_g'X_g u_2$ belong to the domain of attraction of stable laws with an index of stability α.

Remark 3 (Index of stability, tail index, and CLT)**.** Assumption 1 is rather general and covers the most common situations. To see this, note that, when $\alpha \in (0,2)$, Theorem 2.24 of de la Peña, Lai, and Shao (2009) suggests that $v'S_g$ (respectively, $u_1'X_g'X_gu_2$) belongs to the domain of attraction of an α-stable distribution if and only if

$$\mathbb{P}(|v'S_g| > t) = t^{-\alpha}L_1(t), \qquad \big(\text{respectively, } \mathbb{P}(|u_1X_gX_g'u_2| > t) = t^{-\alpha}L_2(t)\big)$$

$$\lim_{t\to\infty}\frac{\mathbb{P}(v'S_g > t)}{\mathbb{P}(|v'S_g| > t)} = p, \quad p \in [0,1], \quad \left(\text{respectively, } \lim_{t\to\infty}\frac{\mathbb{P}(u_1'X_g'X_gu_2 > t)}{\mathbb{P}(|u_1'X_g'X_gu_2| > t)} = \widetilde{p}, \quad \widetilde{p} \in [0,1]\right)$$

for some slowly varying $L_1(\cdot)$ and $L_2(\cdot)$, where $L_2(\cdot)$ and $\widetilde{p}$ may depend on u_1, u_2. The first condition in this alternative characterization requires the tails the cluster specific unit follows a power law with an exponent α, which coincides with the standard definition of the tail index (or the tail exponent) of heavy-tailed random variables in the literature on the extreme value theory (see, e.g. Resnick 2007). See, for example, Theorems 1 and 4 in Sasaki and Wang (2022) for low level sufficient conditions for this index of stability condition, which are framed in terms of cluster size distributions and moments of the regressors and error term. Known as the balancing condition, the second condition in this alternative characterization imposes a mild restriction on the existence of limiting ratios of one-sided tail probabilities over the two-sided tail probability; it rules out some irregular, infinitely oscillating type situations such that these limiting ratios do not exist. This condition only imposes restrictions in the limit and accommodates a wide range of tail behaviors as p (respectively, $\widetilde{p}$) are permitted to be either 0 or 1. When $\alpha < 2$, the variances do not exist and the central limit theorems (CLT) are inapplicable.

On the other hand, when $\alpha = 2$, the limiting α-stable distribution must be normal (cf. Geluk and de Haan, 2000, Theorem 2) and we say the random variables of interest belong to the domain of attraction of the normal distribution. It covers most of the common situations when variance is finite and thus CLT can be applied. It also covers some non-standard cases with a normal limiting distribution but without a finite variance, e.g., a Pareto random variable with a shape parameter (Pareto exponent) of 2.

The next theorem states the main result of this paper focusing on the case of $\alpha \in (1, 2)$ for the moment. We will deal with the remaining case later.

Theorem 1 (Cluster robust inference by score subsampling)**.** *Suppose that Assumption 1 is satisfied for $\alpha \in (1, 2)$. If $b \to \infty$ and $b/G = o(1)$ as $G \to \infty$, then*

$$\sup_{t \in \mathbb{R}} |L^*_{G,b}(t) - J^*(t)| \xrightarrow{p} 0$$

and the limiting distribution $J^*(\cdot)$ is continuous. In addition, if $M \to \infty$, then

$$\sup_{t \in \mathbb{R}} |\widehat{L}_{G,b}(t) - J^*(t)| \xrightarrow{p} 0,$$

and thus

$$\mathbb{P}\left((\widehat{\delta} - \delta)/\widehat{\sigma} \leqslant \widehat{c}_{G,b}(1 - a)\right) \to 1 - a.$$

This theorem formally justifies that the method of inference based on the proposed subsampling procedure is asymptotically valid.

Remark 4 (Heavy-tailed cluster sums)**.** In this theorem, we essentially assume that the tails of the distribution of $\|S_g\|$ and $\|X_g' X_g\|$ both follow a power law with the shape parameter (Pareto exponent) in $(1, 2)$, which implies that the variances of S_g and $(X_g' X_g)$ do not exist. See Remark 3. This is a rather general condition in the sense that the heavy tail can come from the distribution of cluster sizes N_g or the distribution of individuals' (X_{gi}', U_{gi}), or both.

Remark 5 (Impossibility of normal approximation)**.** An inspection of the proof of Theorem 1, combined with Remark 2 in LePage et al. (1981), unveils that, when $\alpha < 2$, the largest cluster has an asymptotically non-negligible influence on the limiting α-stable distribution (see also Sec. 1.4 in Samorodnitsky and Taqqu 1994). For illustration, suppose that the regressor and error distributions are uniformly bounded with variances bounded away from zero. Then this setting essentially translates to

$$\frac{\max_{g=1,\ldots,G}\|S_g\|}{G} \sim \frac{\max_{g=1,\ldots,G} N_g}{G} \gg 0.$$

This is in contrast to the conventional assumption

$$\frac{\max_{g=1,\ldots,G} N_g^2}{G} = o_p(1),$$

which is used in the literature of cluster-robust inference based on the normal approximation.[2]

In addition, a necessary and sufficient condition for the limiting distribution of sums of independent random variables to be normal is the uniform asymptotic negligibility condition, i.e., the largest summand (in absolute value) has an asymptotically negligible contribution to the sum (cf. Davidson, 1994, Theorem 23.13). Thus, it is impossible to derive a theoretically valid normal approximation-based procedure of inference in the presence of non-negligibly large clusters without imposing restrictions on within-cluster dependence.

Remark 6 (On CR standard error estimation)**.** The test statistic we consider is the standard t-statistic used in the literature. Its denominator consists of a CR standard error without imposing a null hypothesis. When $\alpha < 2$, the asymptotic variance does not exist, and nor is this "standard error" consistent but remains random asymptotically. This is similar in spirit to the fixed-b asymptotics (e.g., Kiefer and Vogelsang, 2002) in the literature of long-run variance estimation, although the underlying theory is completely different as the fixed-b asymptotics crucially relies on normal approximation and the functional central limit theorem. Showing that this "standard error" with estimated residuals has negligible impact on the asymptotic distribution requires a completely different proof strategy from the conventional approach of those taken in the proof of Theorem 7.6 in Hansen (2022) for example.

[2] It is assumed in the literature of cluster-robust inference based on the normal approximation that $\frac{\max_{g=1,\ldots,G} N_g^2}{N} = o_p(1)$. When $\mathbb{E}[N_g] = c > 0$ exists, this assumption is equivalent to $\frac{\max_{g=1,\ldots,G} N_g^2}{G} = o_p(1)$.

Proof of Theorem 1. Without loss of generality, suppose that X_{gi} is a scalar and $r = 1$, and hence $\delta = \theta$. The proof is divided into two steps. In the first step, we derive the asymptotic distribution of the self-normalized sums that consist of the linear component of the influence function of the estimator. In the second step, we derive the validity of the proposed subsampling inference procedure.

Step 1. Recall that

$$\widehat{\theta} - \theta = \left(\sum_{g=1}^{G} X_g' X_g \right)^{-1} \sum_{g=1}^{G} S_g.$$

We shall derive the asymptotic distribution for the following self-normalized sums of the linear component $\sum_{g=1}^{G} S_g$:

$$SN_{1G}(\theta) := \frac{\sum_{g=1}^{G} S_g}{\sqrt{\sum_{g=1}^{G} S_g^2}}, \qquad SN_{2G}(\theta) := \frac{\sum_{g=1}^{G} S_g}{\sqrt{\sum_{g=1}^{G} \widehat{S}_g^2}}, \tag{2.1}$$

where $\widehat{S}_g = X_g' \widehat{U}_g$. The asymptotic distribution of a properly re-scaled $(\widehat{\theta} - \theta)$ will then follow straightforwardly from the multiplication of Q^{-1} on both the numerator and the denominator. Since $\alpha \in (1, 2)$, Corollary 1 in LePage et al. (1981) yields

$$SN_{1G}(\theta) \xrightarrow{d} \frac{\sum_{k=1}^{\infty} \{\epsilon_k Z_k - (2p-1)\mathbb{E}[Z_k \mathbb{1}(Z_k < 1)]\}}{\sqrt{\sum_{k=1}^{\infty} Z_k^2}} \tag{2.2}$$

as $G \to \infty$, where

$$p = \lim_{t \to \infty} \frac{\mathbb{P}(S_g > t)}{\mathbb{P}(|S_g| > t)},$$

$Z_k = (E_1 + ... + E_k)^{-1/\alpha}$ for each k, $\{E_k\}_k$ are i.i.d. standard exponential random variables, and $\{\epsilon_k\}_k$ are i.i.d. random variables that take the value of 1 with probability p and -1 with probability $(1-p)$ and are independent of $\{Z_k\}_k$.

We now claim that $SN_{2G}(\theta)$ converges in distribution to the same limiting distribution as (2.2). By Theorems 1 and 1′ in LePage et al. (1981),

$$\left(\frac{1}{A_G}\sum_{g=1}^{G} S_g, \frac{1}{A_G^2}\sum_{g=1}^{G} S_g^2\right) \xrightarrow{d} (S, V) =: \left(\sum_{k=1}^{\infty}\{\epsilon_k Z_k - (2p-1)\mathbb{E}[Z_k \mathbb{1}(Z_k < 1)]\}, \sum_{k=1}^{\infty} Z_k^2\right) = O_p(1) \tag{2.3}$$

holds for $A_G = G^{1/\alpha}L_1(G)$, where Z_k, ϵ_k, and p are defined below Equation (2.2), and $L_1(\cdot)$ is slowly varying at ∞; and

$$\frac{1}{(A'_G)^2}\sum_{g=1}^{G}(X'_g X_g)^2 \xrightarrow{d} \sum_{k=1}^{\infty}\widetilde{Z}_k^2 = O_p(1) \tag{2.4}$$

holds where $A'_G = G^{1/\alpha}L_2(G)$, $\widetilde{Z}_k = (\widetilde{E}_1 + ... + \widetilde{E}_k)^{-1/\alpha}$ for each k, $\{\widetilde{E}_k\}_k$ are i.i.d. standard exponential random variables, and $L_2(\cdot)$ is slowly varying at ∞. Because $\alpha \in (1,2)$ and L_1 is slowly varying at ∞, Equation (2.3) implies the consistency

$$\|\widehat{\theta} - \theta\| = \left\|\left(\sum_{g=1}^{G} X'_g X_g\right)^{-1}\sum_{g=1}^{G} S_g\right\| = O_p(L_1(G)G^{-(1-1/\alpha)}) = o_p(1) \tag{2.5}$$

under Assumption 1. Using $\widehat{U}_g = U_g + X_g(\theta - \widehat{\theta})$ and $\widehat{S}_g = S_g + X'_g X_g(\theta - \widehat{\theta})$, where $\widehat{U}_g = (\widehat{U}_{g1}, ..., \widehat{U}_{gN_g})'$, we can write

$$\begin{aligned}\frac{1}{A_G^2}\sum_{g=1}^{G}\widehat{S}_g^2 =& \frac{1}{A_G^2}\sum_{g=1}^{G} S_g^2 + \frac{1}{A_G^2}\sum_{g=1}^{G}\left(\widehat{S}_g - S_g\right)\widehat{S}_g + \frac{1}{A_G^2}\sum_{g=1}^{G} S_g\left(\widehat{S}_g - S_g\right)\\ =& \frac{1}{A_G^2}\sum_{g=1}^{G} S_g^2 + (1) + (2).\end{aligned}$$

We are going to show that the terms (1) and (2) are $o_p(1)$. First,

$$\begin{aligned}\|(1)\| =& \left\|\frac{1}{A_G^2}\sum_{g=1}^{G}(S_g + X'_g X_g(\theta - \widehat{\theta}))(X'_g X_g(\theta - \widehat{\theta}))'\right\|\\ \leqslant& \left\|\frac{1}{A_G^2}\sum_{g=1}^{G} S_g X'_g X_g\right\|\|\widehat{\theta} - \theta\| + \left\|\frac{1}{A_G^2}\sum_{g=1}^{G}(X'_g X_g)^2\right\|\|\widehat{\theta} - \theta\|^2\end{aligned}$$

$$
\begin{aligned}
&\leqslant \underbrace{\sqrt{\frac{1}{A_G^2}\sum_{g=1}^{G} S_g^2}}_{=O_p(1)} \underbrace{\sqrt{\frac{1}{A_G^2}\sum_{g=1}^{G}(X_g' X_g)^2}}_{=O_p(1)} \underbrace{\|\widehat{\theta}-\theta\|}_{=o_p(1)} + \underbrace{\frac{1}{A_G^2}\sum_{g=1}^{G}(X_g' X_g)^2}_{=O_p(1)} \underbrace{\|\widehat{\theta}-\theta\|^2}_{=o_p(1)} \\
&=o_p(1)
\end{aligned}
$$

holds, where the second inequality follows from the Cauchy-Schwarz inequality and the stochastic orders are due to Equations (2.3), (2.4), and (2.5). Second, similar lines of calculations yield

$$
\|(2)\| = \left\| \frac{1}{A_G^2}\sum_{g=1}^{G} S_g (X_g' X_g(\theta-\widehat{\theta}))' \right\| = o_p(1).
$$

We have now established that

$$
\frac{1}{A_G^2}\sum_{g=1}^{G}\widehat{S}_g^2 = \frac{1}{A_G^2}\sum_{g=1}^{G} S_g^2 + o_p(1),
$$

and consequently, $SN_{1G}(\theta)$ is asymptotically equivalent to $SN_{2G}(\theta)$.

Step 2. We next show the validity of cluster robust score subsampling procedure. Define the conventional subsampling estimator

$$
\breve{\theta}_{b,j} = \left(\sum_{g\in S_j} X_g' X_g\right)^{-1} \sum_{g\in S_j} X_g' Y_g.
$$

Since $B^{-1}-A^{-1} = A^{-1}(A-B)B^{-1}$, we have

$$
\begin{aligned}
\breve{\theta}_{b,j}-\widehat{\theta}_{b,j} &= \left(\sum_{g\in S_j} X_g' X_g\right)^{-1} \sum_{g\in S_j} X_g' Y_g - \left(\frac{G}{b}\right)\left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \sum_{g\in \mathrm{S}_j} X_g' Y_g \\
&= \left(\frac{1}{G}\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\frac{1}{G}\sum_{g=1}^{G} X_g X_g - \frac{1}{b}\sum_{g\in S_j} X_g' X_g\right)\left(\frac{1}{b}\sum_{g\in S_j} X_g' X_g\right)^{-1} \frac{1}{b}\sum_{g\in S_j} X_g' Y_g \\
&=o_p(1)\cdot \breve{\theta}_{b,j}
\end{aligned}
$$

This implies $\widehat{\theta}_{b,j} = \breve{\theta}_{b,j}(1+o_p(1))$. Therefore, in the subsampling process, $\breve{\theta}_{b,j}$ can be replaced by $\widehat{\theta}_{b,j}$ without changing the asymptotic behavior. Thus, it suffices to establish validity of subsampling procedure based on the conventional subsampling estimator $\breve{\theta}_{b,j}$.

Now, since the stable distributions S and V defined in the previous step are both continuous and $V > 0$ with probability 1, $S/V^{1/2}$ is continuously distributed and $J^*(\cdot)$ is continuous. Hence, by invoking Theorem 11.3.1 in Politis et al. (1999), we have

$$\sup_{t \in \mathbb{R}} |L^*_{G,b}(t) - J^*(t)| = o_p(1)$$

as $G \to \infty$, $b \to \infty$, and $b/G = o(1)$. Next, note that $\widehat{L}_{G,b}$ is an empirical CDF consisting of M i.i.d. summands as we randomly sample the subsample clusters with replacement. By Dvoretzky-Kiefer- Wolfowitz inequality, therefore, we have the uniform convergence of the empirical CDF:

$$\sup_{t \in \mathbb{R}} |\widehat{L}_{G,b}(t) - J^*(t)| = o_p(1)$$

as $M \to \infty$ and $G \to \infty$ This concludes the proof. □

In the case that $\|S_g\|$ and $\|X_g' X_g\|$ are in the domain of attraction of the normal law, our proposed inference procedure based on subsampling continues to hold, regardless of whether the variances of S_g and $(X_g' X_g)$ exist or not. The next corollary supports this claim.

Corollary 1. *Under Assumption 1, the conclusion of Theorem 1 continues to hold for $\alpha = 2$.*

Remark 7 (Normal limiting law with and without a finite variance)**.** As mentioned in Remark 3, the case of $\alpha = 2$ includes situations where the limiting distribution follows a normal distribution after proper centering and scaling. Hence, it includes the common situations where variances are finite and thus the CLT is applicable. In addition, it also covers some special circumstances with a normal limiting law but with $\text{Var}(S_g) = \infty$ and thus CLT is not applicable. Theorem 1 and Corollary 1 together comprehensively cover both heavy- and thin-tailed distributions.

Proof of Corollary 1. The proof is similar to the proof of Theorem 1 with the following minor modifications. First, when $\alpha = 2$, S_g is in the domain of attraction of the normal distribution and hence Theorem 3.4 in Giné et al. (1997) yields

$$SN_{1G}(\theta) \xrightarrow{d} N(0,1).$$

Second, to show the asymptotic equivalence of $SN_1(\theta)$ and $SN_2(\theta)$, note that both S_g and $(X_g'X_g)$ belong to the domain of attraction of the normal law when $\alpha = 2$. We branch into two cases. In case that both S_g and $(X_g'X_g)$ have finite variances, we have

$$\frac{1}{G}\sum_{g=1}^{G}\widehat{S}_g^2 = \frac{1}{G}\sum_{g=1}^{G}S_g^2 + o_p(1) \xrightarrow{p} \operatorname{Var}(S_g)$$

by following the standard argument for consistency of the CR variance estimator. In case their variances do not exist, Lemma 3.1 in Giné et al. (1997) yields

$$\frac{1}{A_G^2}\sum_{g=1}^{G}S_g^2 \xrightarrow{p} 1$$

for A_G such that

$$\frac{1}{A_G}\sum_{g=1}^{G}(S_g - \mathbb{E}[S_g]) \xrightarrow{d} N(0,1).$$

A similar argument holds when S_g is replaced by $(X_g'X_g)$. Then, the arguments for bounding $\|(1)\|$ and $\|(2)\|$ in the proof of Theorem 1 still go through, and thus for the self-normalized sums defined in Equation (2.1), it holds that $SN_2(\theta) = SN_1(\theta) + o_p(1)$. Finally, for the validity of the subsampling procedure, we now invoke Theorem 2.2.1 in Politis et al. (1999) and note that the limiting distribution is normal and hence continuous. □

In situations with $\alpha \in (1,2)$, one might hope that the use of the self-normalized CLT can restore the asymptotic normality. Section 4.3 in Sasaki and Wang (2022) shows that, when $\alpha < 2$, there exists a counterexample that the asymptotic distribution is non-gaussian even when a self-normalized test statistic is employed. It turns out that a much stronger conclusion can be made, that is, the score being in the domain of attraction of the normal distribution is not only sufficient, but also necessary for the limiting distribution of the t-statistic to be normal. Therefore, when $\alpha < 2$, it is impossible to derive a valid unconditional inference procedure based on normal approximation without further imposing conditions on within cluster dependence.

Corollary 2. *Suppose that Assumption 1 is satisfied for an* $\alpha \in (1, 2]$*, then the t-statistic* $(\widehat{\delta} - \delta)/\widehat{\sigma}$ *is asymptotically normal if and only if* $\alpha = 2$*.*

Proof. The if part of the statement follows from the proof of Corollary 1. The only if part is a direct implication of Theorem 3.4 in Giné et al. (1997) and the fact that for any $\alpha \in (1, 2]$, the self-normalized sums defined in Equation (2.1) satisfy $SN_2(\theta) = SN_1(\theta) + o_p(1)$, as shown in the proofs for Theorem 1 and Corollary 1. □

2.2. **Choosing the Number of Subsample Clusters.** For the choice of b in practice, we adapt the minimum volatility method (Algorithm 9.3.3) from Section 9.3.2 in Politis et al. (1999) to our framework of cluster-robust inference. For subsampling to be valid, b needs to grow with the number G of clusters but at a slower rate. If b is too close to G, then all the subsampled t-statistics will be almost identical to the full-sample t-statistic, resulting in a subsampling distribution being too tight and thus in under-coverage by confidence intervals. On the other hand, if b is too small, then the subsampled t-statistics will be noisy and can result in either under-coverage or over-coverage. Thus, intuitively, we wish to select a b that is in a stable range for the test statistic. The following algorithm formalizes such an idea.

Algorithm 1 (Minimum volatility method for cluster-robust inference)**.**

(1) For $b \in \{b_{small}, b_{small} + 1, ..., b_{big}\}$*, compute the critical value* $\widehat{c}_{G,b}(1 - a)$ *at a desired significance level* a*.*

(2) For $b \in \{b_{small} + k, b_{small} + k + 1, ..., b_{big} - k\}$*, compute a volatility index* VI_b *of the critical value, i.e., the standard deviation of the values* $\widehat{c}_{G,b-k}(1 - a), ..., \widehat{c}_{G,b}(1 - a), ..., \widehat{c}_{G,b+k}(1 - a)$ *for a small positive integer* k*.*

(3) Pick b^* that has the smallest VI_{b*} and the corresponding confidence interval.

Remark 8. As pointed out in Section 11.5 in Romano and Wolf (1999), as empirical bootstrap is not valid in the presence of heavy-tailed distributions, the common calibration method for the choice of subsampling block size cannot be used in our setup.

3. Inconsistency of the Bootstraps

It is well-known that empirical bootstrap is inconsistent when the variance of the summand is infinite (c.f. Athreya 1987; Knight 1989). It follows straightforwardly that the pairs cluster bootstrap (Cameron, Gelbach, and Miller, 2008) is inconsistent in the context with heavy-tailed distributed cluster sizes. The wild cluster bootstrap (Cameron et al., 2008) is a popular alternative resampling method in the CR context, which has been shown in various simulation studies to be well-behaved under thin-tailed distributed cluster size setups. Validity of the wild cluster bootstrap in the case of thin-tailed cluster size has been shown in, e.g. Djogbenou, MacKinnon, and Nielsen (2019) under fairly general conditions. As their proof relies crucially on Lyapunov's CLT, however, their arguments do not hold in the presence of heavy-tailed cluster sizes (see Remark 5). A remaining and potentially more interesting question is whether one can prove its validity using an alternative argument. The following result suggests that such efforts are ill-fated when $\alpha < 2$. For simplicity of illustration, consider the case of a univariate regression with only the intercept, i.e. a cluster sampled mean $\widehat{\theta} = N^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} Y_{gi}$ with the cluster specific population mean normalized to $\theta = \mathbb{E}\left[\sum_{i=1}^{N_g} Y_{gi}\right] = 0$ without loss of generality. Suppose that the parameter for inference is θ. Under the null hypothesis that $\theta = 0$, the standard cluster robust t-statistic can be formed as

$$T_G = \frac{\sum_{g=1}^{G} \sum_{i=1}^{N_g} Y_{gi}}{\sqrt{\sum_{g=1}^{G} \left(\sum_{i=1}^{N_g} (Y_{gi} - \widehat{\theta}) \right)^2}}.$$

The wild-cluster-bootstrap version of the estimator is defined by $\widehat{\theta}^* = N^{-1} \sum_{g=1}^{G} v_g^* \sum_{i=1}^{N_g} Y_{gi}$, where $(v_g^*)_{g=1}^{G}$ are i.i.d. Rademacher auxiliary random variables generated by the researcher independently from the observed data. The null-imposed wild cluster bootstrap test statistic is defined by

$$T_G^* = \frac{\sum_{g=1}^{G} v_g^* \sum_{i=1}^{N_g} Y_{gi}}{\sqrt{\sum_{g=1}^{G} \left(v_g^* \sum_{i=1}^{N_g} (Y_{gi} - \widehat{\theta}^*) \right)^2}}.$$

Denote $Y_{1:G} = (Y_{gi} : g = 1, ..., G, i = 1, ..., N_g)$. As Theorem 1 implies continuity of the limiting distribution of T_G, the wild cluster bootstrap is consistent if, as $G \to \infty$,

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(T_G^* \leqslant t | Y_{1:G}) - \mathbb{P}(T_G \leqslant t)| = o_p(1).$$

Theorem 2. *Under the above setup and Assumption 1, if $\alpha \in (1, 2)$, then the wild cluster bootstrap with Rademacher auxiliary r.v.'s is inconsistent.*

Proof of Theorem 2. Write

$$T_G = \frac{S_G}{\sqrt{V_G}} := \frac{A_G^{-1} \sum_{g=1}^{G} \left(\sum_{i=1}^{N_g} Y_{gi} \right)}{\sqrt{A_G^{-2} \sum_{g=1}^{G} \left(\sum_{i=1}^{N_g} (Y_{gi} - \widehat{\theta}) \right)^2}} \qquad \text{and}$$

$$T_G^* = \frac{S_G^*}{\sqrt{V_G^*}} := \frac{A_G^{-1} \sum_{g=1}^{G} v_g^* \left(\sum_{i=1}^{N_g} Y_{gi} \right)}{\sqrt{A_G^{-2} \sum_{g=1}^{G} \left(v_g^* \sum_{i=1}^{N_g} (Y_{gi} - \widehat{\theta}^*) \right)^2}}.$$

Let $\mathbb{P}$ denote the probability measure for the data and $\mathbb{P}^*$ denote the probability measure of Rademacher auxiliary random variables. Define

$$p = \lim_{t \to \infty} \frac{\mathbb{P}\left(\sum_{i=1}^{N_g} Y_{gi} > t \right)}{\mathbb{P}\left(\left| \sum_{i=1}^{N_g} Y_{gi} \right| > t \right)}.$$

Write $W_g = \left| \sum_{i=1}^{N_g} Y_{gi} \right|$ and the order statistics of $W_1, ..., W_G$ as follows:

$$W_{G1} \geqslant W_{G2} \geqslant ... \geqslant W_{GG}.$$

The rescaled counterpart is denoted by $Z_{Gg} = A_G^{-1} W_{Gg}$, for $g = 1, ..., G$ – recall that $A_G = G^{1/\alpha} L(G)$ for a slow varying $L(\cdot)$ is defined right before Assumption 1. For each G, we can collect them into a countably long vector

$$Z^G = (Z_{G1}, ..., Z_{GG}, 0, 0, ...) \in \mathbb{R}^\infty.$$

Similarly defined is the countably long sign vector

$$\epsilon^G = (\epsilon_{G1}, ..., \epsilon_{GG}, 1, 1, ...) \in \mathbb{R}^\infty,$$

where ϵ_{Gg} indicates the sign such that $\sum_{i=1}^{N_h} Y_{hi} = \epsilon_{Gg} W_{Gg}$ for the cluster h that corresponds to the g-th order statistic W_{Gg} for each $g = 1, ..., G$, $G \in \mathbb{N}$. By Lemmas 1 and 2 in LePage et al. (1981), we have

$$Z^G \xrightarrow{d} Z = (Z_1, Z_2, ...) \quad \text{and} \quad \epsilon^G \xrightarrow{d} \epsilon = (\epsilon_1, \epsilon_2, ...),$$

where $\{Z_k\}_k$ and $\{\epsilon_k\}$ are defined in the proof for Theorem 1. In addition, since $\mathbb{R}^\infty$ is a complete separable metric space under the metric

$$d((x_1, x_2, ...), (y_1, y_2, ...)) = \sum_{k=1}^{\infty} \frac{1}{2^k} \cdot \frac{|x_k - y_k|}{1 + |x_k - y_k|},$$

following Skorohod's representation theorem, on an adequately chosen probability space,

$$d(Z^G, Z) \to 0 \quad \text{and} \quad d(\epsilon^G, \epsilon) \to 0$$

$\mathbb{P}$-almost surely. Denote the countable vector of i.i.d. Rademacher random variables by $v^* = (v_1^*, v_2^*, ...) \in \mathbb{R}^\infty$, which is invariant of G. We now claim that the weak convergence

$$S_G^* = \sum_{g=1}^{G} \epsilon_{Gg} Z_{Gg} v_g^* \xrightarrow{d^*} S^* := \sum_{k=1}^{\infty} \epsilon_k Z_k v_k^*$$

for (Z, ϵ) with $\mathbb{P}$-probability one, where the convergence in distribution $\xrightarrow{d^*}$ is with respect to $\mathbb{P}^*$. Note that the limiting random variable on the right-hand side is well-defined since

$$\mathbb{E}^* \left[\epsilon_k Z_k v_k^*\right] = 0 \text{ for all } k \text{ and}$$

$$\sum_{k=1}^{\infty} \mathbb{E}^* \left[(\epsilon_k Z_k v_k^*)^2 \right] = \sum_{k=1}^{\infty} Z_k^2 < \infty$$

$\mathbb{P}$-almost surely. The convergence in distribution is shown following the same arguments as in the proof of Theorem 2 in Knight (1989) with i.i.d. Rademacher random variables v_k^* in place of their centered i.i.d. Poisson random variables $(M_k^* - 1)$. Specifically, observe that $Z_k \to 0$ as $k \to \infty$ $\mathbb{P}$-almost surely. Following Equation (12) in the proof of Theorem 1 in LePage et al. (1981), define $\mathcal{Z} \subset \mathbb{R}^{\infty}$ be the subspace consists of countable sequences $z = (z_1, z_2, ...)$ such that $z_1 \geqslant z_2 \geqslant ... \geqslant 0$ (note that $\mathcal{Z}$ is also a complete separable space with the inherited topology). Subsequently, for a fixed $\varepsilon > 0$, define $\phi : \mathcal{Z} \times \{-1, 1\}^{\infty} \times \{-1, 1\}^{\infty}$ by

$$\phi(z, \epsilon, v^*) = \begin{cases} \sum_{k=1}^{\infty} \epsilon_k z_k \mathbb{1}(z_k > \epsilon) v_k^* & \text{if } z_k \to 0 \text{ as } k \to \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Then ϕ is a continuous mapping with respect to the product topology. Thus by the continuous mapping theorem as well as the convergences of $d(Z^G, Z) \to 0$ and $d(\epsilon^G, \epsilon) \to 0$ with $\mathbb{P}$-probability one established earlier, for any $\varepsilon > 0$,

$$\sum_{g=1}^{G} \epsilon_{Gg} Z_{Gg} \mathbb{1}(Z_{Gg} > \varepsilon) v_g^* \xrightarrow{d^*} \sum_{k=1}^{\infty} \epsilon_k Z_k \mathbb{1}(Z_k > \varepsilon) v_k^*$$

for (Z, ϵ) with $\mathbb{P}$-probability one. In addition, note that

$$\mathbb{E}^* \left[\left(\sum_{g=1}^{G} \epsilon_{Gg} Z_{Gg} \mathbb{1}(Z_{Gg} \leqslant \varepsilon) v_g^* \right)^2 \right] = \sum_{g=1}^{G} Z_{Gg}^2 \mathbb{1}(Z_{Gg} \leqslant \varepsilon) \mathrm{Var}^*(v_k^*) \leqslant \sum_{k=1}^{\infty} Z_k^2 \mathbb{1}(Z_k \leqslant \varepsilon)$$

holds almost surely in $\mathbb{P}$ and the right-hand side converges to zero as $\varepsilon \to 0$, which implies via Markov's inequality that, for any $\delta > 0$,

$$\lim_{\varepsilon \to 0} \limsup_{G \to \infty} \mathbb{P}^* \left(\left| \sum_{k=1}^{\infty} \epsilon_{Gk} Z_{Gk} \mathbb{1}(Z_{Gk} \leqslant \varepsilon) v_k^* \right| > \delta \right) = 0$$

$\mathbb{P}$-almost surely. Finally, for any $\delta > 0$,

$$\lim_{\varepsilon \to 0} \mathbb{P}^* \left(\left| \sum_{k=1}^{\infty} \epsilon_k Z_k \mathbb{1}(Z_k \leqslant \varepsilon) v_k^* \right| > \delta \right) = 0$$

$\mathbb{P}$-almost surely, which follows immediately from the fact that

$$\mathbb{E}^*\left[\left(\sum_{k=1}^{\infty}\epsilon_k Z_k \mathbb{1}(Z_k \leqslant \varepsilon)v_k^*\right)^2\right] = \sum_{k=1}^{\infty} Z_k^2 \mathbb{1}(Z_k \leqslant \varepsilon) \to 0$$

$\mathbb{P}$-almost surely as $\varepsilon \to 0$. Combining these results yields that

$$S_G^* \xrightarrow{d^*} S^* = \sum_{k=1}^{\infty}\epsilon_k Z_k v_k^*$$

for (Z, ϵ) with $\mathbb{P}$-probability one. On the other hand, recall from Step 1 in the proof of Theorem 1 that

$$S_G = \sum_{g=1}^{G}\epsilon_{Gg}Z_{Gg} \xrightarrow{d} S := \sum_{k=1}^{\infty}\{\epsilon_k Z_k - (2p-1)\mathbb{E}[Z_k \mathbb{1}(Z_k \leqslant 1)]\},$$

by Theorem 1 in LePage et al. (1981). Note that Z_k, ϵ_k, and v_k^* are all mutually independent from each other. Therefore, the limiting distribution of S_G^* given $Y_{1:G}$, i.e. S^* conditionally on (Z, ϵ), differs from, S, the limiting α-stable distribution of S_G with positive $\mathbb{P}$-probability.

Next, to cope with the denominator term of S_G^*, note that, combined with the law of large numbers, the above weak convergence of S_G^* also implies

$$\begin{aligned}
\widehat{\theta}^* =& \frac{1}{N}\sum_{g=1}^{G}\epsilon_{Gg}W_{Gg}v_g^* \\
=& \frac{1}{c+o_p(1)} \cdot \frac{1}{G}\sum_{g=1}^{G}\epsilon_{Gg}W_{Gg}v_g^* \\
=& \underbrace{\frac{1}{c+o_p(1)}}_{=O_p(1)} \cdot \underbrace{\frac{A_G}{G}}_{=\frac{L(G)}{G^{1-1/\alpha}}} \cdot \underbrace{\sum_{g=1}^{G}\epsilon_{Gg}Z_{Gg}v_g^*}_{=O_p(1)} = o_p(1).
\end{aligned}$$

Thus, the denominator term, $(V_G^*)^{1/2}$, of S_G^* turns out to be asymptotically independent of the auxiliary Rademacher random variables v_g^*:

$$V_G^* = \frac{1}{A_G^2}\sum_{g=1}^{G}\left(v_g^*\sum_{i=1}^{N_g}(Y_{gi}+o_p(1))\right)^2 = \sum_{g=1}^{G}Z_{Gg}^2 + o_p(1).$$

Given $Y_{1:G}$, the denominator is asymptotically constant. Following Step 1 in the proof of Theorem 1, we have

$$V_G = \sum_{g=1}^{G} Z_{Gg}^2 + o_p(1) \xrightarrow{d} \sum_{k=1}^{\infty} Z_k^2 = O_p(1).$$

Thus, given $Y_{1:G}$, the denominator term $(V_G^*)^{1/2}$ is a fixed value, while the original limit of the denominator is an $(\alpha/2)$-stable, non-degenerate continuous distribution. Hence, the limiting distribution of V_G^* given $Y_{1:G}$ and the unconditional limiting distribution of V_G differs with non-zero $\mathbb{P}$-probability.

Finally, note that $V_G^* > 0$ $\mathbb{P}$-almost surely. Thus, the fact that $(S_G^*, V_G^*) \xrightarrow{d^*} \left(\sum_{k=1}^{\infty} \epsilon_k Z_k v_k^*, \sum_{k=1}^{\infty} Z_k^2\right)$ for almost every (Z, ϵ) and the continuous mapping theorem yields that

$$T_G^* \xrightarrow{d^*} \frac{\sum_{k=1}^{\infty} \epsilon_k Z_k v_k^*}{\sqrt{\sum_{k=1}^{\infty} Z_k^2}}$$

for (Z, ϵ) with $\mathbb{P}$-probability one. This, together with the unconditional limiting distribution of T_G implies the conclusion that the unconditional limiting distribution of T_G and the conditional limiting distribution of T_G^* differs with positive $\mathbb{P}$-probability. The inconsistency then follows.

□

References

Arcones, M. A. and E. Giné (1989): "The bootstrap of the mean with arbitrary bootstrap sample size," in *Annales de l'IHP Probabilités et Statistiques*, vol. 25, 457–481.

Athreya, K. (1987): "Bootstrap of the mean in the infinite variance case," *Annals of Statistics*, 724–731.

Bugni, F., I. Canay, A. Shaikh, and M. Tabord-Meehan (2022): "Inference for cluster randomized experiments with non-ignorable cluster sizes," *arXiv preprint arXiv:2204.08356*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90, 414–427.

Cameron, A. C. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 50, 317–372.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*, OUP Oxford.

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

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393–412.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997): *Modelling Extremal Events: for Insurance and Finance*, Springer Science & Business Media.

Feller, W. (1971): *An Introduction to Probability Theory and Its Applications*, vol. 2, New York: John Wiley,.

Geluk, J. and L. de Haan (2000): "Stable probability distributions and their domains of attraction: a direct approach," *Probability and Mathematical Statistics*, 20, 169–188.

Giné, E., F. Götze, and D. M. Mason (1997): "When is the Student t-statistic asymptotically standard normal?" *Annals of Probability*, 25, 1514–1531.

Hansen, B. (2022): *Econometrics*, Princeton University Press.

Kiefer, N. M. and T. J. Vogelsang (2002): "Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation," *Econometrica*, 70, 2093–2095.

Knight, K. (1989): "On the bootstrap of the sample mean in the infinite variance case," *Annals of Statistics*, 1168–1175.

Kojevnikov, D. and K. Song (2023): "Some Impossibility Results for Inference With Cluster Dependence with Large Clusters," *Journal of Econometrics*, Forthcoming.

LePage, R., M. Woodroofe, and J. Zinn (1981): "Convergence to a stable distribution via order statistics," *Annals of Probability*, 9, 624–632.

Logan, B. F., C. Mallows, S. Rice, and L. A. Shepp (1973): "Limit distributions of self-normalized sums," *Annals of Probability*, 1, 788–809.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272–299.

Politis, D. N. and J. P. Romano (1994): "Large sample confidence regions based on subsamples under minimal assumptions," *Annals of Statistics*, 22, 2031–2050.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer Science & Business Media.

Resnick, S. (1987): *Extreme Values, Regular Variation and Point Processes*, Springer-Verlag.

Resnick, S. I. (2007): *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer Science & Business Media.

Romano, J. P. and M. Wolf (1999): "Subsampling inference for the mean in the heavy-tailed case," *Metrika*, 50, 55–69.

Samorodnitsky, G. and M. Taqqu (1994): *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, vol. 1, CRC Press.

Sasaki, Y. and Y. Wang (2022): "Non-robustness of the cluster-robust inference: with a proposal of a new robust method," *arXiv preprint arXiv:2210.16991*.

Zolotarev, V. M. (1986): *One-Dimensional Stable Distributions*, vol. 65, American Mathematical Society.

(H. D. Chiang) Department of Economics, University of Wisconsin-Madison, William H. Sewell Social Science Building, 1180 Observatory Drive, Madison, WI 53706, USA.

Email address: hdchiang@wisc.edu

(Y. Sasaki) Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA.

Email address: yuya.sasaki@vanderbilt.edu

(Y. Wang) Department of Economics, Syracuse University, 200 Eggers Hall, Syracuse, NY 13244-1020, USA.

Email address: ywang402@syr.edu