Genuinely Robust Inference for Clustered Data[*]

Harold D. Chiang[†] Yuya Sasaki[‡] Yulong Wang[§]

Abstract

Conventional cluster-robust inference can be invalid when data contain clusters of unignorably large size. We formalize this issue by deriving a necessary and sufficient condition for its validity, and show that this condition is frequently violated in practice: specifications from 77% of empirical research articles in *American Economic Review* and *Econometrica* during 2020–2021 appear not to meet it. To address this limitation, we propose a genuinely robust inference procedure based on a new cluster score bootstrap. We establish its validity and size control across broad classes of data-generating processes where conventional methods break down. Simulation studies corroborate our theoretical findings, and empirical applications illustrate that employing the proposed method can substantially alter conventional statistical conclusions.

Keywords: cluster-robust inference, cluster score bootstrap, unignorably large cluster, domain of attraction, extreme value theory

JEL Code: C12, C18, C46

[*] First arXiv date: August 23, 2023. This version: September 22, 2025. We owe special thanks to A. Colin Cameron, Tetsuya Kaji, and Kevin Song for their invaluable comments that significantly improved our methodology. We also benefited from discussions with Ivan Canay, Bruce Hansen, Michael Jansson, Seojeong (Jay) Lee, Michael Leung, James MacKinnon, Francesca Molinari, Ulrich Müller, Jack Porter, Frank Schorfheide, Xiaoxia Shi, Max Tabord-Meehan, Tim Vogelsang, and seminar participants at numerous institutions and conferences. We thank Hong Xu and Siyuan Xu for their excellent research assistance. All remaining errors are ours.

[†] Assistant Professor of Economics, University of Wisconsin-Madison. hdchiang@wisc.edu

[‡] Brian and Charlotte Grove Chair and Professor of Economics, Vanderbilt University. Email: yuya.sasaki@vanderbilt.edu

[§] Associate Professor of Economics, Syracuse University. Email: ywang402@syr.edu

1 Introduction

Cluster-robust (CR) standard errors are designed to account for within-cluster correlations. Such correlations often arise by construction, for example, within an industry (Hersch, 1998) or within a state (Bertrand, Duflo, and Mullainathan, 2004). Today, even when a model does not inherently induce cluster dependence, the application of CR methods using observable group identifiers has become a common practice.

The foundational theory (White, 1984; Liang and Zeger, 1986; Arellano, 1987) for CR inference methods assumes small cluster sizes N_g (uniformly bounded above by $\overline{N} < \infty$) with a large number of clusters, $G \to \infty$, where N_g denotes the number of entities in the g-th cluster for $g \in \{1, 2, \ldots, G\}$. Procedures based on this theory are implemented through the '`cluster()`' and '`vce(cluster)`' options in Stata, and they are utilized in nearly all, if not all, empirical studies that report CR standard errors.

It has been recognized that large cluster sizes N_g can result in inflated CR standard errors (e.g., Cameron and Miller, 2015, p. 324). Recent theoretical advancements (Carter, Schnepel, and Steigerwald, 2017; Djogbenou, MacKinnon, and Nielsen, 2019; Hansen and Lee, 2019; Hansen, 2022b; Bugni, Canay, Shaikh, and Tabord-Meehan, 2024) accommodate larger cluster sizes N_g, eliminating the requirement that $N_g \leqslant \overline{N}$ and thereby broadening the applicability of the '`cluster()`' and '`vce(cluster)`' options, among others. With this said, they still impose the restriction $\max_g N_g^2/N \to 0$ of vanishing maximum cluster size relative to the square root of the whole sample size $N = \sum_{g=1}^{G} N_g$ as $G \to \infty$.

A natural question is whether the relaxed condition $\max_g N_g^2/N \to 0$ accommodates a wide range of data sets. To answer this, we analyze empirical papers published in top journals.[1] All of these articles employ the aforementioned Stata options for CR standard errors, thereby implicitly assuming $\max_g N_g^2/N \to 0$. Table 1 summarizes the number of articles with $\max_g N_g^2/N$ falling into each bin on a logarithmic scale. Notably, 55 percent (respectively, 39, 29, and 16 percent) of the articles use data sets where $\max_g N_g^2/N \geqslant 1$

[1] We studied all articles published in *American Economic Review* and *Econometrica* between 2020 and 2021. Among them, we extracted a list of papers reporting estimation and inference results based on regressions, IV regressions, and their variants. Furthermore, we focus on articles using publicly available data sets for replication. See Section 3 for further details of this study.

The Empirical Distribution of $\max_g N_g^2/N$ in Empirical Economic Research: 2020–2021

$\max_g N_g^2/N$		0.1–	1–	10–	100–	⩾1000
	<0.1	1	10	100	1000	
American Economic Review	4	8	4	1	3	1
Econometrica	2	0	1	2	1	4
Total	6	8	5	3	4	5
	(19%)	(26%)	(15%)	(10%)	(13%)	(16%)

Table 1: Distribution of articles by the value of $\max_g N_g^2/N$, classified into the bins $[0, 0.1)$, $[0.1, 1)$, $[1, 10)$, $[10, 100)$, $[100, 1000)$, and $[1000, \infty)$ on a logarithmic scale. The sample consists of papers published in the *American Economic Review* and *Econometrica* during 2020–2021 that report cluster-robust standard errors for regression or IV regression using publicly available replication data sets. For papers reporting multiple regressions, we record the largest value of $\max_g N_g^2/N$ across specifications.

(respectively, $\geqslant 10$, $\geqslant 100$, and $\geqslant 1000$). In other words, the condition $\max_g N_g^2/N \to 0$, required for the validity of conventional CR inference, may not hold for a nontrivial portion of these published articles.

The condition $\max_g N_g^2/N \to 0$ is sufficient but not necessary for asymptotic normality, implying that the adequacy of normality-based confidence intervals and tests cannot be evaluated solely by assessing the plausibility of this condition. To address this, we establish a necessary and sufficient condition for the validity of conventional cluster-robust (CR) inference – see Theorem 1. Specifically, the limiting distribution is normal if and only if the score of the largest cluster is ignorable. When clusters are *unignorably large*, regression estimates exhibit non-Gaussian limiting distributions, as illustrated in Figure 1.[2] Using this characterization, formal statistical tests based on Sasaki and Wang (2023) reject the null hypothesis of normality in 24 of the 31 papers (77 percent) reported in Table 1–see Table 2.

Non-Gaussian limiting distributions invalidate conventional critical values, such as "1.96," as well as bootstrap critical values. For example, using 1.96 results in sizes of 0.053, 0.087, and 0.250 (instead of the desired 0.050) when the nuisance parameter α equals 1.75, 1.50, and 1.25, respectively, as shown in Figure 1. The empirical bootstrap fails in these cases of infinite variance, and the widely used wild cluster bootstrap and pairs cluster bootstrap are also inconsistent. Later, we formally establish these negative results as Proposition 1.

[2] Details on these non-Gaussian distributions are provided in Section 3.

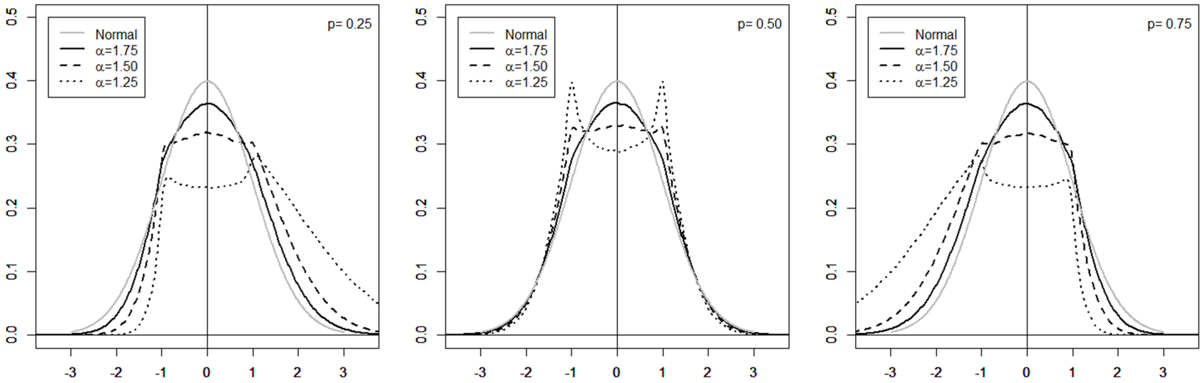

Figure 1: Illustration of non-Gaussian limiting distributions arising from unignorably large clusters. The different distributional shapes, indexed by α and p, are described in Section 3.

To address this issue, we introduce the cluster score (CS) bootstrap, a novel inferential method for clustered data that extends the m-out-of-n and score bootstraps. This method provides valid critical values adaptively across all limiting distributions depicted in Figure 1. In addition, we provide a data-driven choice of the tuning parameter and justify its theoretical validity.

Relation to the Literature: The literature on cluster-robust inference has a long history, dating back to White (1984), Liang and Zeger (1986), and Arellano (1987). For a comprehensive review, we refer readers to Cameron and Miller (2015, 2025) and MacKinnon, Nielsen, and Webb (2023). More recently, sampling frameworks in which cluster sizes are treated as random variables have been investigated by Bai, Liu, Shaikh, and Tabord-Meehan (2022), Bugni et al. (2024), and Cavaliere, Mikosch, Rahbek, and Vilandt (2024). We adopt a model-based perspective with an increasing number of clusters and unrestricted intra-cluster dependence, or, asymptotically equivalently, a sampling-based perspective where the growing number of sampled clusters represents a negligible fraction of the superpopulation. This framework is well suited to the empirical contexts encountered in most applications.[3]

[3] An alternative framework assumes a fixed number of clusters with growing cluster sizes, where asymptotic normality can be derived under additional assumptions of weak intra-cluster dependence, as in Hansen (2007), Ibragimov and Müller (2010), Canay, Santos, and Shaikh (2021). Ibragimov and Müller (2016) and Hansen (2022c) consider inference under gaussian assumptions. Another line of research advances the integration of design-based and sampling-based asymptotics, particularly under explicit treatment assignment schemes such as randomized experiments, as considered by Abadie, Athey, Imbens, and Wooldridge (2023). Extending our method to this design-based framework is an avenue for future research.

In an insightful recent work, Kojevnikov and Song (2023) establish an impossibility result for consistent estimation of the asymptotic variance when the sample contains a single large cluster under a triangular array setup. They further provide a necessary and sufficient condition on the cluster structure for the asymptotic variance to be consistently estimable. Our findings complement their result by showing that normal approximation for t-statistics fails in the presence of unignorably large clusters. Furthermore, our proposed procedure overcomes this limitation, as it does not rely on consistent variance estimation. We demonstrate that the self-normalized statistic converges in distribution and formally derive its limiting stable distribution in such settings. Importantly, the implementation of the bootstrap inference procedure does not require knowledge of the unknown rate or consistent variance estimation, owing to the self-normalizing nature of the test statistics.

The aspect of our paper that establishes non-Gaussianity under certain conditions connects to a branch of the econometrics literature exploring the possibility of non-Gaussian limiting distributions and, in some cases, impossibility results. For instance, Hirano and Porter (2012) show that regular estimation, and hence Gaussian asymptotics, are impossible for a class of estimators characterized by the maximum. More directly related, Menzel (2021) highlight the potential non-Gaussianity of estimators under two-way clustering and establish an impossibility result on the uniform consistency of tests. In contrast, we demonstrate that non-Gaussianity can arise even under one-way clustering. Moreover, such negative results may be even more pervasive in this widely used empirical setting.

Our key distributional approximation results build on Logan, Mallows, Rice, and Shepp (1973), LePage, Woodroofe, and Zinn (1981), and Giné, Götze, and Mason (1997). For theoretical foundations of probability and statistics with heavy-tailed distributions, we refer readers to Resnick (1987, 2007) and Samorodnitsky and Taqqu (1994). Also see Ibragimov, Ibragimov, and Walden (2015); Chernozhukov, Fernández-Val, and Kaji (2016, 2017) for applications in economics and finance. Our inference procedure builds on the resampling theory developed in Arcones and Giné (1989, 1991) and Bickel and Sakov (2008). For the related discussions on the inconsistency of empirical bootstrap for means of random variables with infinite variance, see, e.g., Athreya (1987) and Knight (1989).

2 The Model

While the idea extends to a general class of econometric models, we consider the linear model[4]

$$Y_{gi} = X_{gi}'\theta + U_{gi}, \qquad \mathbb{E}[U_{gi}|X_g] = 0 \quad i = 1, ..., N_g,$$

for ease of exposition as well as its popular use in practice, where $X_g = (X_{g1}, \ldots, X_{gN_g})'$, $U_g = (U_{g1}, \ldots, U_{gN_g})'$, $g \in \{1, \ldots, G\}$ indexes clusters, and N_g denotes the size of the g-th cluster. Define the OLS estimator and its cluster-robust (CR) variance estimator by

$$\widehat{\theta} = \left(\sum_{g=1}^{G}\sum_{i=1}^{N_g} X_{gi}X_{gi}'\right)^{-1} \sum_{g=1}^{G}\sum_{i=1}^{N_g} X_{gi}Y_{gi} = \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} \sum_{g=1}^{G}(X_g'X_g\theta + S_g) \quad \text{and} \tag{2.1}$$

$$\widehat{V}^{\text{CR}} = a_G \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} \left(\sum_{g=1}^{G} \widehat{S}_g\widehat{S}_g'\right) \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1}, \tag{2.2}$$

respectively, for some finite sample adjustment factor a_G such that $a_G \to 1$ as $G \to \infty$, 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 of writing, we set $a_G = 1$ throughout as it does not affect our asymptotic arguments.

Consider a linear transformation $\delta = r'\theta$ of the regression coefficient vector θ, such that $r \in \mathbb{R}^{\dim(\theta)}$ and $\|r\| = 1$, as the parameter of interest. Let the corresponding estimator and its CR standard error be denoted by

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

respectively. We are interested in conducting inference for δ using the t-statistic

$$\frac{(\widehat{\delta} - \delta)}{\widehat{\sigma}} = \frac{r'(\widehat{\theta} - \theta)}{\sqrt{r'\left(\sum_{g=1}^{G} X_g'X_g\right)^{-1}\left(\sum_{g=1}^{G} \widehat{S}_g\widehat{S}_g'\right)\left(\sum_{g=1}^{G} X_g'X_g\right)^{-1} r}} \tag{2.3}$$

[4] The assumption $\mathbb{E}[U_{gi}|X_g] = 0$, while standard in the literature, is stronger than required for our asymptotic results. It can be relaxed to $\mathbb{E}[\sum_{i=1}^{N_g} X_{gi}U_{gi}] = 0$.

based on the CR standard error.

To state our assumption, we introduce a few definitions. 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

$$\frac{\sum_{g=1}^{G} \xi_g - D_G}{A_G} \xrightarrow{d} \eta \qquad \text{as } G \to \infty.$$

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 ∞ (cf. Proposition 2.2.13 in Embrechts, Klüppelberg, and Mikosch 1997). If $\alpha \in (1, 2]$, then D_G can be chosen to be $G \cdot \mathbb{E}[\xi_g]$. 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*. Although this concept may look esoteric to some readers, it essentially states that a sum of i.i.d. random variables, after being suitably centered and normalized, converges in distribution to a limiting random variable, and it, in particular, encompasses the standard cases where central limit theorems (CLTs) hold. In other words, econometricians and economists adopting the standard inference (e.g., the conventional critical value of 1.96) implicitly make this (and even stronger) assumption.

Assumption 1. $(X_g'X_g, S_g)_{g=1}^G$ are i.i.d., follow a non-degenerate distribution, $\mathbb{E}[N_g] = c \in (0, \infty)$, and the design matrix satisfies $G^{-1}\sum_{g=1}^G X_g'X_g = Q + o_p(1)$ for a finite positive definite matrix Q. 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 $\alpha \in (1, 2]$.

This assumption is arguably general. It is significantly weaker than requiring a central limit theorem to hold and even covers scenarios where the asymptotic normality fails. It also encompasses two notable cases frequently considered in economics and econometrics.

First, the case of $\alpha = 2$ encompasses the conventional setting in which $r'(\widehat{\theta} - \theta)$ attains the standard convergence rate of $\sqrt{G}$ by the central limit theorem. In this case, the limiting α-stable distribution is necessarily Gaussian (cf. Geluk and de Haan, 2000, Theorem 2). This scenario also includes certain non-standard cases with a Gaussian limiting distribution but without finite variance, such as a Pareto random variable with Pareto exponent equal to 2. The vast majority of econometric papers deriving asymptotic normality implicitly rely on this high-level condition in Assumption 1, or on even stronger ones.

Second, the case of $\alpha < 2$ entails a power law (de la Peña, Lai, and Shao, 2009, Theorem 2.24), i.e.,

$$P(|v'S_g| > t) = t^{-\alpha}L_1(t), \qquad P(|u_1X_g'X_gu_2| > t) = t^{-\alpha}L_2(t), \tag{2.4}$$

for some slowly varying functions $L_1(\cdot)$ and $L_2(\cdot)$, where $L_2(\cdot)$ may depend on u_1 and u_2. In this case, the index α of stability coincides with the Pareto exponent[5] β, in the sense that $\alpha = \min\{\beta, 2\}$. Thus, when $\alpha < 2$, the score has infinite variance. For more precise details, see Theorem 4 in Appendix A.1. In this case, *unignorably large* clusters are indeed unignorable, since the sample sum of the (scaled) scores becomes asymptotically proportional to the (scaled) score of the largest cluster; see Remark 7 in Appendix A.2 for further discussion. Hence, the asymptotic distribution cannot be Gaussian in this case.

The literature in urban economics and economic geography establishes that (truncated) city size distributions frequently exhibit a power-law behavior in the upper tail, with estimated exponents typically in the range of 1 to 1.5; see Eeckhout (2004); Ioannides and Skouras (2013) and references therein. Consequently, when observations are strongly correlated at the city level, this implies $\alpha < 2$, yielding a non-Gaussian limiting distribution.

The i.i.d. requirement across clusters in Assumption 1 is standard in this literature (e.g., Bugni et al. 2024; Cavaliere et al. 2024; Bai et al. 2022). This assumption is mild because: (1) the conditional distributions of S_g and $X_g'X_g$ given $N_g = n_g$ may vary across n_g; and (2) the distributions of individuals within each cluster need not be identical. Moreover, S_g and X_g may be arbitrarily correlated with the cluster size N_g, provided that the regression exogeneity condition is satisfied.

To simplify the exposition, we focus on the case where $v'S_g$ and $u_1'X_g'X_gu_2$ share a common stability index α. This simplification is rationalized if the tail behavior of their distributions is driven by the tail behavior of the cluster-size distribution N_g; see Section 3.1 for

[5] Specifically, the Pareto distribution has CDF $F(t) = 1 - t^{-\beta}$ for $t \geqslant 1$.

an illustrative example. However, this simplification is adopted only for notational simplicity and can be relaxed at the cost of substantially more cumbersome exposition.

3 Fragility of the Conventional CR Methods

This section shows that conventional methods of cluster-robust (CR) inference are valid if and only if $\alpha = 2$. In other words, they necessarily fail when $\alpha < 2$. We begin with heuristic discussions in Section 3.1 and then develop formal results in Section 3.2. We further report how frequently cases with $\alpha < 2$ arise in empirical economic applications.

3.1 Heuristic Discussions

The intuition behind the fragility of conventional CR methods is as follows. When $\alpha < 2$, the cluster size N_g does not have a finite variance. If intra-cluster dependence is non-trivial, this infinite variance of N_g is inherited by the score S_g, causing the CLT for OLS (and also other estimators) to fail. Cases with $\alpha < 2$ are quite plausible in empirical data, and we show that this is indeed the case for the majority of recent empirical papers published in *Econometrica* and the *American Economic Review*, as discussed in Section 3.2 in detail.

To provide a simple and transparent illustration, consider the sample average

$$\widehat{\theta} = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{N_g} Y_{gi},$$

which is a special case of the OLS estimator (2.1) with $X_{gi} = 1$. The true parameter is the mean $\theta = \mathbb{E}[Y_{gi}]$, which is normalized to $\theta = 0$ without loss of generality. For clarity, suppose the extreme case of perfect intra-cluster dependence, i.e., $Y_{gi} \equiv Y_g$ for all $i \in \{1, \ldots, N_g\}$ within each cluster g. For simplicity, also assume that N_g is independent of Y_g. These assumptions are made purely for expositional clarity and are not essential for our results.

In this case, we obtain

$$\sqrt{N}\widehat{\theta} = \frac{G^{-1/2} \sum_{g=1}^{G} N_g Y_g}{\sqrt{\frac{1}{G} \sum_{g=1}^{G} N_g}}.$$

The denominator converges to $\sqrt{\mathbb{E}[N_g]}$ as long as $\alpha > 1$. For the numerator, note that $\mathrm{Var}\Big(G^{-1/2}\sum_{g=1}^{G} N_g Y_g\Big) = \mathrm{Var}[N_g]\cdot\mathrm{Var}[Y_g]$, which is infinite when $\alpha < 2$. Indeed, Theorem 1 of Geluk and de Haan (2000) implies that, if the distribution of $N_g Y_g$ is α-stable, then the limiting distribution

$$x \;\mapsto\; \lim_{G\to\infty} \Pr\left(\frac{1}{a_G}\sum_{g=1}^{G} N_g Y_g - b_G > x\right),$$

for some sequences $a_G \simeq G^{1/\alpha} \to \infty$ and $b_G \in \mathbb{R}$, has the characteristic function

$$\psi_\alpha(s) = \exp\left\{-\Big(|s|^\alpha + is(1-\alpha)\tan(\alpha\pi/2)\tfrac{|s|^{\alpha-1}-1}{\alpha-1}\Big)\right\},$$

which differs from the Gaussian characteristic function. Thus, the CLT for $\widehat{\theta}$ fails. Further discussion of this example is provided in Appendix A.4.

In summary, the stability index α determines both the convergence rate and the limiting distribution. When intra-cluster correlation is non-trivial, the tail heaviness of N_g carries over to that of S_g. Consequently, the t-ratio of the conventional CR method is asymptotically normal *if and only if* $\alpha = 2$. While we currently present heuristic arguments in a simplified setting, we formalize and generalize this claim in Theorem 1 and Proposition 1 below.

Remark 1 (Bias from Trimming Large Clusters)**.** In practice, researchers may trim large clusters with $N_g > k$ for some threshold k. While such trimming may appear to mitigate problems arising from non-Gaussian limiting distributions induced by unignorably large clusters, it introduces bias and thereby undermines the validity of inference. Consider

$$0 = \mathbb{E}\left[\sum_{i=1}^{N_g} Y_{gi}\right] = \mathbb{E}[kY_g\,\mathbb{1}\{N_g \leqslant k\}] + \mathbb{E}[(N_g - k)Y_g\,\mathbb{1}\{N_g > k\}] =: \theta(k) + \lambda(k).$$

Here, $\theta(k)$ represents the estimand of the trimmed procedure, while $\lambda(k)$ denotes the associated bias term. This bias $\lambda(k)$ is generally nonzero whenever the distribution of Y_g depends on N_g. Hence, naively trimming large clusters can result in invalid inference. ▲

3.2 Formal Theory

The following theorem formalizes and generalizes the discussion from the previous subsection.

Theorem 1 (Necessary and sufficient condition)**.** *If Assumption 1 is satisfied for an* $\alpha \in (1, 2]$*, then the t-statistic* (2.3) *is asymptotically normal if and only if* $\alpha = 2$*.*

A proof is provided in Appendix C.3.

The theorem implies that conventional CR inference based on common variance estimators, such as CR1, CR2, CR3, and the jackknife, together with normal critical values (e.g., ≈ 1.96 for the 97.5th percentile) fails whenever $\alpha < 2$.

With Theorem 1, we now characterize the curves shown in Figure 1 from Section 1. The left, middle, and right panels of Figure 1 display the limiting distributions of the t-statistic under $p = 0.25$, 0.50, and 0.75, respectively, where

$$p = \lim_{t\to\infty} \frac{P(v'S_g > t)}{P(|v'S_g| > t)}, \tag{3.1}$$

for v as given in Assumption 1, measures the limiting asymmetry of tail probabilities. Each panel in Figure 1 depicts three non-Gaussian limiting distributions corresponding to $\alpha =$ 1.25, 1.50, and 1.75 with distinct line styles, together with the normal reference case ($\alpha =$ 2.00). The key takeaway is that conventional methods of CR inference, which rely on normal approximation, become increasingly size-distorted as α decreases and as p deviates from 0.5.

Another class of conventional approaches consists of cluster bootstraps. Two main bootstrap-based CR inference methods are commonly used in the literature: the pairs cluster bootstrap and the wild cluster bootstrap (Cameron, Gelbach, and Miller, 2008). It is well established that the empirical bootstrap is inconsistent when the variance of the score is infinite (cf. Athreya, 1987; Knight, 1989). In light of the power-law characterization (2.4), the pairs cluster bootstrap, essentially the empirical bootstrap applied to cluster-wise sums treated as independent units, is inconsistent under Assumption 1 with $\alpha < 2$. Moreover, Theorem 5 in Appendix A.3 shows that the wild cluster bootstrap is likewise inconsistent under Assumption 1 with $\alpha < 2$. The following proposition summarizes these results.

Proposition 1 (Failure of the Conventional Cluster Bootstraps)**.** *If Assumption 1 is satisfied for an* $\alpha < 2$*, then the pairs cluster bootstrap and the wild bootstrap methods are both inconsistent.*

Given that the case of $\alpha < 2$ invalidates all conventional methods of CR inference, a natural question is how frequently such cases arise in empirical economics. To address this, we examined all articles published in two leading journals, the *American Economic Review* and *Econometrica*, during 2020–2021. From these, we extracted the subset of papers reporting estimation and inference results based on regressions, IV regressions, and related variants. We further restricted attention to articles that employ publicly available datasets due to replicability.

For these articles, we test the null hypothesis $H_0 : \alpha = 2$ against the alternative $H_1 : \alpha < 2$ for the score. Such a test can be implemented via the likelihood ratio test of Sasaki and Wang (2023), which considers the surrogate null hypothesis $H_0 : \beta \geqslant 2$ against the alternative $H_1 : \beta < 2$ in light of (2.4), where β denotes the tail exponent of the score.[6]

Table 2 summarizes the set of papers included in our study. The first two columns report the journals and years of publication. The next column, "All #," indicates the total number of eligible articles according to the selection criteria described above. The column group labeled "Cluster" contains articles in which CR inference is applied to at least one regression result. Within this group, the column "#" reports the number of such articles, while the column "Test $\alpha < 2$" reports the fraction of these articles for which the test rejects the null hypothesis in one or more regression specifications. The final row presents the column totals.

During 2020–2021, the *American Economic Review* published 30 articles that met our selection criteria. Of these, 21 reported CR standard errors. The null hypothesis is rejected in 16 of these 21 papers. In other words, inference based on the conventional CR method may be misleading in approximately 76% of the articles employing it.

During 2020–2021, *Econometrica* published 14 articles that met our selection criteria. Of these, 10 reported CR standard errors. The null hypothesis is rejected in 8 of these 10 papers. In other words, inference based on the conventional CR method may be misleading in approximately 80% of the articles employing it.

Combining the two journals, we find that inference may be misleading in as many as

[6] The test of the null hypothesis $H_0 : \beta \geqslant 2$ against the alternative $H_1 : \beta < 2$ is implemented using the Stata command `testout y x1 x2 ..., cluster(cid)` for least-squares estimation and `testout y x1 x2 ..., iv(z) cluster(cid)` for instrumental variables estimation, both following Sasaki and Wang (2023).

	Year of	All	Cluster		
Journal	Publication	#	#	Test $\alpha < 2$	
American Economic Review	2020	15	10	7/10	(70%)
American Economic Review	2021	15	11	9/11	(82%)
	Subtotal	30	21	16/21	(76%)
Econometrica	2020	12	7	7/8	(88%)
Econometrica	2021	3	2	1/2	(50%)
	Subtotal	15	10	8/10	(80%)
	Total	45	31	24/31	(77%)

Table 2: The column "All – #" reports the total number of eligible articles that use regressions or IV regressions with publicly available replication data. The column "Cluster – #" reports the number of these articles that employ cluster-robust (CR) inference. The column "Cluster – Test $\alpha < 2$" reports the rejection rate of the null hypothesis $\alpha = 2$ among articles using CR inference. The tests of $\alpha = 2$ against the alternative $\alpha < 2$ are implemented with the Stata commands "`testout y x1 x2 ..., cluster(cid)`" for regressions and "`testout y x1 x2 ..., iv(z) cluster(cid)`" for IV regressions, following Sasaki and Wang (2023).

77% of the 31 articles that employ the conventional CR method. Thus, problematic practice appears to be prevalent even in these highly influential outlets.[7] All of the above issues with conventional CR methods motivate our proposed approach: the cluster score (CS) bootstrap, which accommodates non-Gaussian limiting distributions, to be presented in Section 4.

4 The Cluster Score Bootstrap

In light of the limitations of conventional CR inference methods discussed in the previous section, we introduce a novel cluster score (CS) bootstrap procedure to approximate the limiting distribution of $(\widehat{\delta}-\delta)/\widehat{\sigma}$. This procedure remains valid whether the limiting distribution is Gaussian or non-Gaussian. Section 4.1 describes the proposed method, and Section 4.2 provides its theoretical justification.

[7] Spreadsheets of all the test results with specific papers and specific equations are available upon request.

4.1 The Method

Our objective is to conduct statistical inference for δ using the t-statistic defined in (2.3). Let the CDF J_G^* of the t-statistic be

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

We will show that, under suitable regularity conditions, J_G^* converges to the CDF J^* of the corresponding limiting distribution.

Let b denote the number of resampled clusters, chosen according to Algorithm 1 (to be presented in Section 4.2). For a large positive integer M, draw M i.i.d. multinomial random vectors $(w_1^j, \ldots, w_G^j)_{j=1}^M$, each with b trials and uniform cell probability $1/G$, independently of the data. This is equivalent to sampling b clusters with replacement uniformly from the G clusters in the data, and w_g^j records the counts of how many times the g-th cluster is selected in the j-th bootstrap sample. Define the CS bootstrap estimator and its associated variance estimator by

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

where $\widehat{S}_{g,j} = X_g'\big(Y_g - X_g\widehat{\theta}_{b,j}\big)$.

Note that the inverse factor $\left(\sum_{g=1}^{G} X_g' X_g\right)^{-1}$ is computed from the full unweighted sample, whereas the linear component and its variance are constructed from the bootstrap sample. Practical motivations for this feature will be discussed in Remark 3.

Define the bootstrapped empirical distribution function $\widehat{L}_{G,b}$ of $(\widehat{\delta}_{b,j} - \widehat{\delta})/\widehat{\sigma}_{b,j}$ by

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

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

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

As will be formally established in Section 4.2, this critical value is guaranteed to satisfy

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

as $G \rightarrow \infty$. Hence, the CS bootstrap method provides asymptotically valid inference.

Practical Implication: For the t-statistic, one may continue to use the conventional CR "standard error" $\widehat{\sigma}$ even though it may diverge.[8] However, rather than relying on conventional Gaussian critical values (e.g., $\Phi^{-1}(0.025) \approx -1.96$ and $\Phi^{-1}(0.975) \approx 1.96$), one should instead employ the bootstrap-based critical values $\widehat{c}_{G,b}(0.025)$ and $\widehat{c}_{G,b}(0.975)$ obtained from the CS bootstrap procedure. These critical values, for example, can be used to construct a 95% confidence interval for δ. ▲

Remark 2 (Practicality of the method)**.** Even though the convergence rate of $\widehat{\delta}-\delta$ is unknown, our inference remains valid because it is based on a self-normalized statistic. Moreover, our procedure does *not* require estimation of the unknown stability index α, nor is it necessary to estimate the slowly varying functions L_1 and L_2. These features represent important practical advantages of our proposed method, as these nuisance parameter estimation problems are well known to be challenging in the statistics literature. ▲

Remark 3 (Finite sample non-invertibility of other resampling methods)**.** Compared with conventional resampling methods, the CS bootstrap offers two advantages. First, because it does not require recomputation of the inverse factor at each bootstrap iteration, the method is computationally more efficient. Second, and more importantly, in finite samples, when the regressors include a cluster-specific binary treatment variable or other dummies that are highly correlated within a cluster, the matrix $\sum_{g=1}^{G} w_g^j X_g' X_g$ is often singular for small b, as is common in cluster-RCT settings. Consequently, the resampled OLS estimator may

[8] Note that the "standard error" $\widehat{\sigma}$ does not converge in probability when $\alpha < 2$.

be undefined in a non-negligible fraction of bootstrap iterations. This problem also arises in other cluster-based resampling methods, such as the jackknife, subsampling, and the conventional bootstrap. In practice, several *ad hoc* “fixes,” such as employing a generalized inverse or dropping such realizations, are often used, though their theoretical justification remains unclear. By contrast, the proposed CS bootstrap procedure, which relies on the full unweighted matrix $\sum_{g=1}^{G} X_g' X_g$, avoids this issue in a theoretically supported manner. ▲

Remark 4 (Inference using parametric bootstrap)**.** The t-statistic has a complicated but well-defined class of limiting distributions, as illustrated in Figure 1. A natural alternative approach is to bootstrap critical values from this known limiting distribution for inference, as suggested in Cornea-Madeira and Davidson (2015). However, this requires estimation of the unknown parameters: the index of stability α and the measure of limiting symmetry p as defined in (3.1). Our simulation results show that inference based on estimated values of α and p performs poorly in finite sample. Moreover, estimating α and p requires selecting tuning parameters that are inherently *ad hoc* choices. The resulting inference is highly sensitive to this tuning and remains imprecise unless the number of clusters is quite large (e.g., exceeding 2000). For these reasons, we do not recommend this parametric bootstrap-based approach in our setup. ▲

Remark 5 (Subsampling)**.** If the i.i.d. count vectors $(w_1^j, \ldots, w_G^j)$ are instead generated by sampling b out of G units without replacement, the procedure would entail a version of score subsampling. Theoretical results for this alternative subsampling approach are established in a manner similar to those for the CS bootstrap. ▲

4.2 Theoretical Properties

We now provide theoretical support for our proposed CS bootstrap method. The following theorem provides a formal justification of its robust asymptotic validity.

Theorem 2 (Cluster-Robust Inference by the CS Bootstrap)**.** *Suppose that Assumption 1 is satisfied. If $b \to \infty$ and $b/G = o(1)$ as $G \to \infty$, and $M \to \infty$, then*

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

for a continuous limiting distribution $J^*(\cdot)$. Thus, for any significance level $a \in (0,1)$,

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

The proof is non-trivial and proceeds by considering two distinct cases. The first, corresponding to $\alpha < 2$, is formalized in Lemma 1 in Appendix A.2 and proved in detail in Appendix C.1. The second case, $\alpha = 2$, is analyzed in Appendix C.2, which also synthesizes the two regimes to establish Theorem 2. Here, asymptotics are taken with respect to $G \to \infty$ for a given DGP; further studies on uniformity can be found in Appendix B.

While the theory requires $b \to \infty$ and $b/G = o(1)$ as $G \to \infty$, in practice the researcher must select a finite value of b. This choice should be neither too large nor too small. Intuitively, if b is chosen too close to G, the largest clusters are sampled too frequently in the bootstrapped t-statistics, which prevents the procedure from adequately reflecting the heavy-tailed nature of the DGP when $\alpha < 2$. Conversely, if b is too small, the bootstrapped t-statistics become excessively noisy. Thus, one seeks a value of b that lies in a stable range, such that small perturbations of b (e.g., increasing or decreasing it by one) have only minimal impact on the bootstrap distribution. In the context of the m-out-of-n bootstrap, Bickel and Sakov (2008) formalized this idea and proposed a data-driven algorithm with theoretical guarantees for its validity. Here, we introduce a modified version of their algorithm tailored to our setting.

Algorithm 1 (Data-Driven Choice of b)**.**

(i) *Let $b_\ell = \lceil q^\ell \cdot \phi(G) \rceil$ for each $\ell \in \mathbb{N}$, where $\lceil a \rceil$ stands for the smallest integer greater than or equal to a, $q \in (0,1)$, and ϕ is a strictly sub-linear function.*[9]

(ii) *Obtain $\widehat{L}_{G,b_\ell}$ for each $\ell \in \mathbb{N}$ by simulation.*

(iii) *Set*

$$\widehat{b} = \operatorname*{argmin}_{\ell \in \mathbb{N}} \sup_{t \in \mathbb{R}} \left| \widehat{L}_{G,b_\ell}(t) - \widehat{L}_{G,b_{\ell+1}}(t) \right|.$$

[9] For instance, we set $\phi(G) = G^{0.99}$. This specification, together with $q = 0.99$, is used in our numerical studies and empirical application.

(iv) If there is a tie, let $\widehat{b}$ be the largest b_ℓ among them.

The following theorem provides theoretical guarantees for the data-driven choice of $\widehat{b}$ in the CS bootstrap. Specifically, it shows that Theorem 2 continues to hold with our data-driven choice of $\widehat{b}$.

Theorem 3 (Cluster-Robust Inference by the Data-Driven CS Bootstrap)**.** *Suppose that Assumption 1 is satisfied and $b = \widehat{b}$ is chosen according to Algorithm 1. Then, the conclusion of Theorem 2 continues to hold.*

A proof is found in Appendix C.4. Hence, following the selection of $\widehat{b}$, asymptotically valid inference can be carried out using the corresponding critical value $\widehat{c}_{G,\widehat{b}}(1-a)$.

5 Simulation Studies

In this section, we present simulation studies evaluating the finite-sample performance of our proposed method of genuinely robust CR inference, based on the cluster score (CS) bootstrap, in comparison with conventional CR methods.

The data-generating design is defined as follows. We consider the cluster treatment model with individual covariates

$$Y_{gi} = \theta_0 + \theta_1 T_g + \sum_{j=1}^{K} \theta_j X_{g,i,j+1} + U_{gi},$$

following MacKinnon, Nielsen, and Webb (2022, Equation (40)), among others. The binary treatment variable T_g equals one for $\lceil 0.2G \rceil$ clusters and zero for the remaining $G - \lceil 0.2G \rceil$ clusters, where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. Cluster sizes are drawn independently as $N_g \sim \lceil \text{Pareto}(1, \alpha) \rceil$ for $g \in \{1, \ldots, G\}$. For each $g \in \{1, \ldots, G\}$, we independently draw N_g-variate random vectors $(\widetilde{X}_{g1j}, \ldots, \widetilde{X}_{gN_gj})' \sim \mathcal{N}(0, \Omega)$ for $j \in \{1, \ldots, K\}$ and $(\widetilde{U}_{g1}, \ldots, \widetilde{U}_{gN_g})' \sim \mathcal{N}(0, \Omega)$ in the baseline design, where Ω is an $N_g \times N_g$ covariance matrix with $\Omega_{ii} = 1$ for all $i \in \{1, \ldots, N_g\}$ and $\Omega_{ii'} = 1/2$ whenever $i \neq i'$. The controls are constructed as $X_{gij} = 0.2F^{-1}_{\text{Beta}(2,2)} \circ \Phi(\widetilde{X}_{gij})$, where $F_{\text{Beta}(2,2)}$ and Φ denote

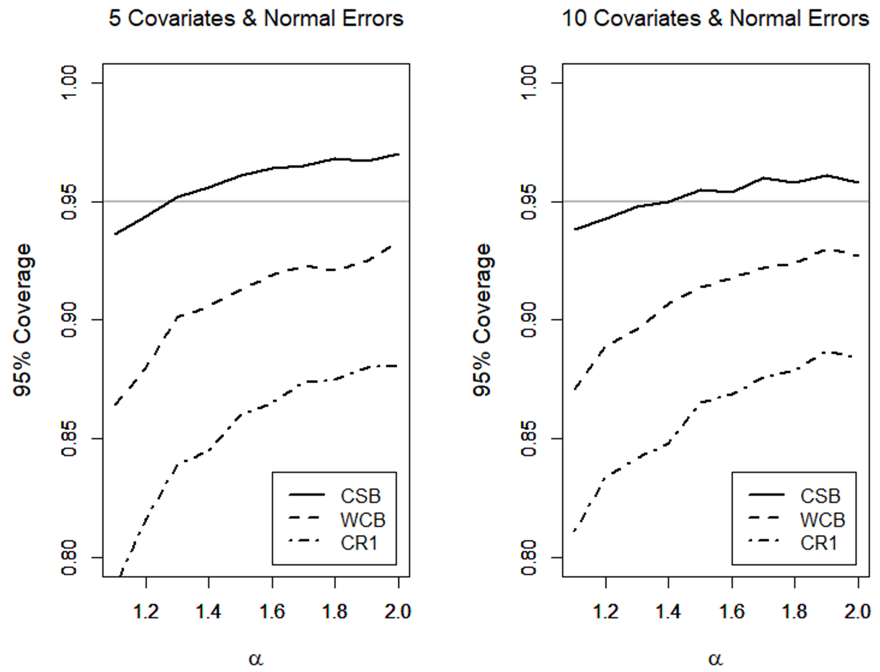

Figure 2: Monte Carlo coverage frequencies for the baseline designs with normal errors. “CSB” (respectively, “WCB” and “CR1”) denotes the CS bootstrap (respectively, wild cluster bootstrap and CR1 standard errors with normal critical values). The nominal 95% coverage probability is indicated by the horizontal gray line.

the CDFs of the Beta$(2, 2)$ and standard normal distributions, respectively. The errors are constructed heteroskedastically as $U_{gi} = 0.2\widetilde{U}_{gi}$ if $T_g = 0$ and $U_{gi} = \widetilde{U}_{gi}$ if $T_g = 1$.

We vary the exponent parameter $\alpha \in \{1.1, 1.2, \ldots, 1.9, 2.0\}$ across simulation sets. The regression coefficients are fixed at $(\theta_0, \theta_1, \theta_2, \ldots, \theta_{K+1})' = (1, 1, 1, \ldots, 1)'$ throughout, while the covariate dimension varies as $K \in \{5, 10\}$. The sample size (i.e., the number of clusters) is set to $G = 50$ across all simulations, which is roughly comparable to the number of U.S. states. Each simulation set consists of 10,000 Monte Carlo iterations.

Figure 2 reports the Monte Carlo coverage frequencies. The horizontal axis represents the value of α, and the vertical axis represents the coverage frequency. In the legend, ‘CSB’ denotes the CS bootstrap, while ‘WCB’ and ‘CR1’ denote the wild cluster bootstrap and the CR1 standard error with normal critical values, respectively. The nominal coverage probability of 95% is indicated by the horizontal gray line at 0.95.

The CS bootstrap performs best, followed by the WCB and the CR1. Overall, the CS bootstrap consistently delivers coverage frequencies closest to the nominal 95% level across the range of α. By contrast, both conventional methods suffer from under-coverage, particularly for small values of α.

6 An Empirical Illustration

Akhtari, Moreira, and Trucco (2022) study the effects of political turnover on various outcomes measuring the quality of public services in Brazil. In their original paper (Table 3, Column 5), they estimate the following linear model by OLS:

$$\text{score}_{gi+1} = \theta_0 + \theta_1 \cdot \mathbb{1}\{\text{IVM}_g < 0\} + \theta_2 \cdot \text{IVM}_g + \theta_3 \cdot \mathbb{1}\{\text{IVM}_g < 0\} \cdot \text{IVM}_g + \theta_4 \cdot \text{score}_{gi} + U_{gi}.$$

The dependent variable, score_{gi+1}, is the test score of fourth-grade students in the year following an election. The main explanatory variable, IVM_g, is the incumbent vote margin. Thus, $\mathbb{1}\{\text{IVM}_g < 0\}$ equals one when the incumbent party loses the election. The parameter of interest is $\delta = \theta_1$, which measures the effect of political turnover on test scores.[10] While the original paper considers alternative 'bandwidths,' we focus on the bandwidth 0.110 to maximize the sample size, following prior work (MacKinnon et al., 2022, Sec. 7.2), which replicates this regression.

The original paper clusters standard errors at the municipality level, and we follow this definition of the cluster unit. There are $G = 2101$ municipalities in the data, with $\max_{1 \leqslant g \leqslant G} N_g^2/N \approx 26$. Thus, the assumption $\max_{1 \leqslant g \leqslant G} N_g^2/N \to 0$, under which conventional CR inference methods are guaranteed to work, is difficult to justify in this application.

Table 3 reports the p-values for $\delta = \theta_1$ based on alternative inference methods. Column HC1 reports the p-value using conventional inference without clustering, i.e., the HC1 standard error. Columns CR1 and WCB report the p-values using conventional CR inference methods, namely the CR1 standard error and the wild cluster bootstrap,[11] respectively,

[10] This effectively implements a sharp regression discontinuity design, although the original paper estimates the effect by OLS using this linear specification.

[11] While there are four variants of WCB (cf. MacKinnon et al., 2022, Sec. 7.2), they yield identical p-values

	HC1	CR1	WCB	CSB
p-value	0.000	0.006	0.006	0.266

Table 3: p-values for the effect $\delta = \theta_1$ of political turnover on fourth-grade students' test scores, based on conventional inference methods (HC1, CR1, WCR) and our proposed CS bootstrap (CSB).

with normal approximation. Finally, column CSB reports the p-value based on our proposed inference method using the CS bootstrap. We employ the same code as in the simulation studies in Section 5, including the choice of b based on the minimum volatility method.

The p-value is zero up to the third digit when standard errors are not clustered (HC1). Conventional CR inference methods (CR1 and WCB) with normal approximation yield larger p-values, but the statistical significance remains unchanged. By contrast, our proposed CS bootstrap method produces a much larger p-value, rendering the effect $\delta = \theta_1$ statistically insignificant, unlike any of the conventional methods. These results highlight that failing to account for potential non-Gaussianity in the limiting distributions, particularly in the presence of unignorably large clusters, can lead to erroneous statistical conclusions.

7 Summary

Conventional methods for cluster-robust inference often fail to provide consistent results in the presence of unignorably large clusters. In this paper, we formalize this limitation by deriving a necessary and sufficient condition for consistency. We document that 77% of empirical research articles published in the *American Economic Review* and *Econometrica* during 2020–2021 contain model specifications fail to satisfy this condition.

To address this challenge, we propose the CS bootstrap and establish its size control across a wide class of data-generating processes where conventional methods break down. Our simulation studies confirm the reliability and effectiveness of the proposed method, underscoring its practical value in overcoming the limitations of existing cluster-robust inference techniques. We further demonstrate the failure of the wild cluster bootstrap in Section A.3

up to the reported digits, and hence we summarize them in a single column.

and discuss the related uniformity issues in Section B in the appendix, reinforcing the need for our proposed approach. Finally, we demonstrate that correctly accounting for potential non-Gaussianity can overturn empirical conclusions. We conclude the paper by modifying a well-known *haiku* by Hirano (1998).

T-stat looks too good.
Use *the cluster score bootstrap*–
significance gone.

Appendix

A Omitted Details

This appendix section collects technical details that are omitted from the main text.

A.1 Alternative Characterization of ξ_g Belonging to the Domain of Attraction of an α-Stable Distribution for $\alpha < 2$

Citing a result from the existing literature, this section presents complete details about the power law characterization (2.4) discussed in Section 2 in the main text.

Theorem 4 (de la Peña et al., 2009, Theorem 2.24)**.** *If $\alpha < 2$, then ξ_g belongs to the domain of attraction of an α-stable distribution if and only if*

$$P(|\xi_g| > t) = t^{-\alpha} L(t) \qquad \textit{and} \tag{A.1}$$

$$\lim_{t \to \infty} \frac{P(\xi_g > t)}{P(|\xi_g| > t)} = p \tag{A.2}$$

for some $p \in [0, 1]$ and some slowly varying function $L(\cdot)$.

The first condition (A.1) means that the tail limit of the absolute value of the random variable of interest has an approximately Pareto tail, or so-called power law. Known as the balancing condition, the second condition (A.2) 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 pathological, 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 are permitted to be even 0 and 1, thereby allowing cases from distributions that are bounded on one side to distributions with heavy two-sided tails.

A.2 Auxiliary Theory Focusing on Cases with $\alpha < 2$

This section presents a lemma that we state and prove on the way to proving Theorem 2 in Section 4.2 in the main text. Namely, for ease of writing, we state our main result focusing on cases with $\alpha \in (1, 2)$. An extension of this result to the general cases with $\alpha \in (1, 2]$ follows as Theorem 2 with additionally accounting for the case with $\alpha = 2$.

Lemma 1. *Suppose that Assumption 1 is satisfied for $\alpha \in (1, 2)$. If $b \to \infty$, $b/G = o(1)$, and $M \to \infty$ as $G \to \infty$, then*

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

and thus

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

A proof is provided in Appendix C.1.

Remark 6 (Heavy-tailed cluster sums)**.** In this lemma, we essentially assume that the tails of the distributions of $\|S_g\|$ and $\|X_g' X_g\|$ both follow the 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 Appendix A.1. This is a rather general condition in the sense that the heavy tail can come from the distribution of cluster sizes N_g, the distribution of individuals' (X_{gi}', U_{gi}), or both. ▲

Remark 7 (Unignorability and impossibility of normal approximation)**.** An inspection of the proof of Lemma 1, combined with Remark 2 in LePage et al. (1981), unveils that, when $\alpha < 2$, the tails of the first component of representation (C.3) satisfies

$$P\left(|\epsilon_1 Z_1 - (2p - 1)\mathbb{E}[Z_1 \mathbb{1}(Z_1 < 1)]| > t\right) \sim P\left(\left|\sum_{k=1}^{\infty} \{\epsilon_k Z_k - (2p - 1)\mathbb{E}[Z_k \mathbb{1}(Z_k < 1)]\}\right| > t\right)$$

as $t \to \infty$. Since the term $|\epsilon_1 Z_1 - (2p - 1)\mathbb{E}[Z_1 \mathbb{1}(Z_1 < 1)]|$ corresponds to the limiting distribution of the absolute value of the scaled score of the largest cluster, it has an asymptotically unignorable influence on the limiting α-stable distribution – see also Section 1.4 in Samorodnitsky and Taqqu (1994). For ease of illustration, suppose that the regressor and error distributions are uniformly bounded and $\mathrm{Cov}(X_{gi}U_{gi}, X_{gi}U_{gi'}|N_g) \geqslant \underline{c} > 0$ for all $i = 1, ..., N_g$ with probability one. This then implies

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

which directly violates the necessary and sufficient condition for the asymptotic variance to be estimable derived in Corollary 4.1 in Kojevnikov and Song (2023), as well as the conventional assumption

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

required in the literature (e.g. Assumption 2 in Hansen and Lee 2019) for normal approximation.[12]

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 unignorably large clusters without imposing restrictions on within-cluster dependence. ▲

Remark 8 (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

[12] It is assumed in the literature of CR 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 $\frac{\max_{g=1,\ldots,G} N_g^2}{G} = o_p(1)$.

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 (2022a). ▲

A.3 Inconsistency of the Wild Cluster Bootstrap under $\alpha < 2$.

The wild cluster bootstrap (Cameron et al., 2008) is a popular alternative resampling method of CR inference. It has been shown in various simulation studies to behave well under $\alpha = 2$. Validity of the wild cluster bootstrap in cases of $\alpha = 2$ has been shown in Djogbenou et al. (2019) under fairly general conditions. As their proof relies crucially on Lyapunov's CLT, however, their arguments do not hold under $\alpha < 2$ – see Remark 7. 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 of inference is θ. Under the null hypothesis $H_0 : \theta = 0$, the standard CR 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 a researcher independently from the observed data $\{\{Y_{gi}\}_{i=1}^{N_g}\}_{g=1}^{G}$. The null-imposed wild cluster bootstrap test statistic is defined by

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

We introduce the notation $Y_{1:G} = (Y_{gi} : g = 1, ..., G, i = 1, ..., N_g)$ for convenience. As

Lemma 1 implies continuity of the limiting distribution of T_G, the wild cluster bootstrap is consistent if

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

Theorem 5 (Inconsistency of the wild cluster bootstrap)**.** *Under the above setup and Assumption 1, if $\alpha \in (1, 2)$, then the wild cluster bootstrap with Rademacher auxiliary random variables is inconsistent.*

A proof can be found in Appendix C.5. Note that the proof strategy for the numerator of the test statistic is closely related to the approach taken in Cavaliere, Georgiev, and Taylor (2013), which demonstrates that a suitably modified wild bootstrap can be valid for symmetrically distributed i.i.d. sample means with infinite variance.

A.4 Details of Section 3.1

Section 3.1 argues that the self-normalized CLT may or may not hold under our framework. This appendix section presents details of this argument.

For the estimand $\theta = \mathbb{E}[Y_{gi}]$ for simplicity, consider the estimator

$$\widehat{\theta} = \frac{1}{N}\sum_{g=1}^{G} S_g,$$

where $S_g = \sum_{i=1}^{N_g} Y_{gi}$ and $N = \sum_{g=1}^{G} N_g$. For simplicity, assume that Y_{gi} is identically distributed with mean zero and variance one, and that Y_{gi} is independent from N_g. Also, assume the cluster-sampling framework in which observations are independent across g. Let Ω_N denote the variance of $\sqrt{N}\widehat{\theta}$, i.e., $\mathbb{E}[N\widehat{\theta}^2]$.

We now consider three cases of within-cluster dependence: (i) Y_{gi} is i.i.d. across i within each g (i.e., no cluster dependence); (ii) $Y_{gi} = Y_{gj}$ for all i and j within the same cluster (i.e., the strongest form of cluster dependence); and (iii) a combination of the cases (i) and (ii).

Case (i) Suppose that Y_{gi} is i.i.d. across i. The self-normalized CLT considers

$$\left(\mathbb{E}[\widehat{\theta}^2]\right)^{-1/2}\widehat{\theta}\xrightarrow{d}\mathcal{N}(0,1).$$

Since $\mathbb{E}[N\widehat{\theta}^2]=\mathbb{E}\left[Y_{gi}^2\right]=1$ under the independence across i and g, we have

$$\left(\mathbb{E}[\widehat{\theta}^2]\right)^{-1/2}\widehat{\theta}=\sqrt{N}\widehat{\theta}=\frac{G^{-1/2}\sum_{g=1}^{G}S_g}{\sqrt{\frac{1}{G}\sum_{g=1}^{G}N_g}}. \tag{A.3}$$

By the law of large numbers and the assumption that N_g is regularly varying with exponent $\alpha>1$, we have

$$\frac{1}{G}\sum_{g=1}^{G}N_g\xrightarrow{d}\mathbb{E}\left[N_g\right]<\infty$$

for the denominator of (A.3). The independence within cluster implies that conditional on $\{N_g\}_{g=1}^{G}$,

$$\frac{1}{\sqrt{N}}\sum_{g=1}^{G}\sum_{i=1}^{N_g}Y_{gi}\xrightarrow{d}\mathcal{N}(0,1).$$

for the numerator of (A.3). Therefore, the self-normalized CLT still holds, but with the convergence rate being $N^{-1/2}$, instead of $G^{-1/2}$ if we treat $\{N_g\}_{g=1}^{G}$ as fixed sequences of constants. Now consider $\{N_g\}_{g=1}^{G}$ as random variables. Given the Pareto tail of N_g, we have that

$$N=\sum_{g=1}^{G}N_g=O_p(G).$$

It follows that $G^{-1/2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}Y_{gi}=O_p(1)$.

Case (ii) Consider the case with perfect within-cluster dependence, i.e., $Y_{gi}\equiv Y_g$ for all $i\in\{1,...,N_g\}$ for each g. In this case, $S_g=\sum_{i=1}^{N_g}Y_{gi}=N_gY_g$, yielding that

$$\sqrt{N}\widehat{\theta}=\frac{G^{-1/2}\sum_{g=1}^{G}N_gY_g}{\sqrt{\frac{1}{G}\sum_{g=1}^{G}N_g}}.$$

The denominator still converges to $\sqrt{\mathbb{E}\left[N_g\right]}$. For the numerator, since N_gY_g is i.i.d. across g and the two factors are independent with regularly varying tails, Mikosch

(1999, Proposition 1.3.9) implies that the product $N_g Y_g$ also has regularly varying tail with exponent $\alpha < 2$. Therefore, $G^{-1/2}\sum_{g=1}^{G} Z_g = G^{-1/2}\sum_{g=1}^{G} N_g Y_g$ is no longer $O_p(1)$. More specifically, $\mathrm{Var}[G^{-1/2}\sum_{g=1}^{G} N_g Y_g]$ is equal to $\mathrm{Var}[N_g] \cdot \mathrm{Var}[Y_g]$, which is infinite given $\alpha < 2$. In fact, Geluk and de Haan (2000, Theorem 1) implies that if the distribution of $N_g Y_g$ is α-stable, under some sequences of constants $a_G \simeq n^{1/\alpha} \to \infty$ and $b_G \in \mathbb{R}$, the limiting distribution

$$\lim_{G\to\infty} P\left(\frac{1}{a_G}\sum_{g=1}^{G} S_g - b_G > x\right)$$

has the characteristic function

$$\psi_\alpha(s) = \exp\left\{-\left(|s|^\alpha + is(1-\alpha)\tan(\alpha\pi/2)\frac{|s|^{\alpha-1}-1}{\alpha-1}\right)\right\}.$$

Thus, the CLT fails, and the asymptotic distribution will be non-Gaussian. Therefore, even the jackknife standard error fails in this scenario. See, for example, Figures 5 and 6 in MacKinnon et al. (2022).

Case (iii) Combining the above two cases, we now consider

$$Y_{gi} = \rho_G R_g + U_{gi},$$

where R_g can be thought as a cluster-specific random effect and U_{gi} is a random noise, which is i.i.d. across both i and g. The normalizing constant ρ_G determines the weights of R_g in Y_{gi}. Under this setting, we have

$$\begin{aligned}\sqrt{N}\widehat{\theta} =& \frac{G^{-1/2}\sum_{g=1}^{G} S_g}{\sqrt{\frac{1}{G}\sum_{g=1}^{G} N_g}} \\ =& \frac{G^{-1/2}\rho_G\sum_{g=1}^{G} N_g R_g}{\sqrt{\frac{1}{G}\sum_{g=1}^{G} N_g}} + \frac{G^{-1/2}\sum_{g=1}^{G}\sum_{i=1}^{N_g} U_{gi}}{\sqrt{\frac{1}{G}\sum_{g=1}^{G} N_g}}.\end{aligned} \tag{A.4}$$

Following the same arguments as those in Case (ii), the first item above is asymptotically non-Gaussian (after some suitable normalization), but the second term is asymptotically normal. The orders of magnitudes of them depend on the distribution of (R_g, N_g, U_{gi}). For example, if $\mathbb{E}[R_g] = 0$ and $\mathbb{E}\left[R_g^2\right] < \infty$, then $N_g R_g$ again has a regularly varying tail with exponent $\alpha < 2$ (e.g., Embrechts and Goldie, 1980, Theorem 3). The generalized central limit theorem (e.g., Ibe, 2013, Chapter 11) implies that $\sum_{g=1}^{G} N_g R_g \simeq_p G^{1/\alpha}$. For the second term in (A.4), Case (i) derives that $G^{-1/2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} U_{gi} = O_p(1)$. The non-Gaussian part then dominates the normal part if $\rho_G G^{1/\alpha - 1/2} \to \infty$ as $G \to \infty$. Since $\alpha < 2$, a constant ρ_G will satisfy this condition.

As a final remark, we note that Assumption 3 in Djogbenou et al. (2019) could relax the condition on N_g into that $\max_{1 \leqslant g \leqslant G} N_g/N \to 0$ when the within-cluster dependence is strong. The stochastic counterpart of this assumption fails under our framework where N_g is treated as a random variable. More specifically, consider Case (ii) again for illustration. Let μ_N denote the reciprocal of the variance of $\widehat{\theta}$ conditional on $\{N_g\}_{g=1}^{G}$ as in Djogbenou et al. (2019). The above derivation yields

$$\begin{aligned} \mathrm{Var}[\widehat{\theta} | \{N_g\}_{g=1}^{G}] &= \frac{\sum_{g=1}^{G} N_g^2 \mathrm{Var}[Y_{gi}]}{(\sum_{g=1}^{G} N_g)^2} \\ &= \frac{\sum_{g=1}^{G} N_g^2 \mathrm{Var}[Y_{gi}] G^{-2}}{(G^{-1} \sum_{g=1}^{G} N_g)^2} \\ &\simeq_p G^{2/\alpha - 2}, \end{aligned}$$

and hence $\mu_N \simeq_p G^{2-2/\alpha}$. Therefore, for any constant $\lambda > 0$, we have

$$\mu_N^{\frac{2+\lambda}{2+2\lambda}} \frac{\sup_g N_g}{N} \simeq_p G^{\rho(\lambda)},$$

where $\rho(\lambda) = (2 - 2/\alpha)[(2+\lambda)/(2+2\lambda) - 1/2]$. Recall $\alpha \in (1, 2)$, yielding that $\rho(\lambda) > 0$ for all $\lambda > 0$. Then, the above term diverges with probability approaching one.

B Discussions on Uniformity

In the main text, all asymptotic properties are derived under a fixed data-generating process (DGP). This appendix extends the analysis to uniform size control over a class of DGPs.

Without uniformity, for any given G (regardless of its magnitude), there may exist a DGP P_G such that the rejection probability under the null fails to converge to the nominal level. Uniformity therefore guarantees more reliable inference in finite samples, particularly when G is moderate. To simplify the notations and assumptions, we focus on inference for the mean of a scalar random variable in the current subsection.

Consider a triangular array setup: for each $G \in \mathbb{N}$, suppose that we have an i.i.d. sequence $(S_g)_{g=1}^G = (S_{g,G})_{g=1}^G$, whose distribution is now $P = P_G$. Recall that

$$(\widehat{\delta} - \delta) = \frac{1}{G}\sum_{g=1}^{G} S_g \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{G}\sum_{g=1}^{G} \widehat{S}_g^2,$$

where $\widehat{S}_g = S_g - G^{-1}\sum_{g=1}^G S_g$. The test statistic of interest is again the t-ratio $(\widehat{\delta} - \delta)/\widehat{\sigma}$. Henceforth, we will let $\mathbb{E}_P[\cdot]$ denote the expectation with respect to the DGP, P, if we are to emphasize such a dependence. For any $\varepsilon \in [0, 1)$, define $\mathbf{P}_1(\varepsilon)$ as the set of all the DGPs, P, such that there exist some $p \in [0, 1]$ and $\alpha \in [1 + \varepsilon, 2)$ such that

$$\lim_{t\to\infty} \frac{P(S_g > t)}{P(|S_g| > t)} = p, \quad \text{and} \tag{B.1}$$

$$P(|S_g| > t) = t^{-\alpha} L_P(t) \quad \text{as } t \to \infty \tag{B.2}$$

for an $L_P(\cdot)$ slowly varying at ∞ that can depend on $P = P_G$, and $\mathbb{E}_P[S_g] = 0$ when it is defined. In addition, define $\mathbf{P}_2$ as the set of all DGPs satisfying $\mathbb{E}_P[S_g] = 0$ and the following uniform integrability condition

$$\lim_{\lambda\to\infty} \sup_{P\in\mathbf{P}_2} \mathbb{E}_P\left[\frac{|S_g - \mathbb{E}_P[S_g]|^2}{\sigma^2(P)} \mathbb{1}\left\{\frac{|S_g - \mathbb{E}_P[S_g]|}{\sigma(P)} > \lambda\right\}\right] = 0,$$

where $\sigma^2(P) = \mathbb{E}_P[S_g^2]$ is finite. Finally, define $\mathbf{P}(\varepsilon) = \mathbf{P}_1(\varepsilon) \cup \mathbf{P}_2$. The first set $\mathbf{P}_1(\varepsilon)$ covers the DGPs with heavy tail distributions and with regularly varying tail probabilities so that the variances of S_g are infinite. The second set $\mathbf{P}_2$ covers a rich subset of DGPs in which the variances of S_g are always finite and contains, in particular, the set of DGPs with $2 + \epsilon$ moments for any $\epsilon > 0$. It rules out certain examples such as those in the classical Bahadur-Savage example under which the t-test fails its size control for every sample size;

see Romano (2004) for more details.

First, we note that when $\alpha = 1$, the t-ratio does not converge in distribution in general, except in very special situations. The following is a direct implication of Logan et al. (1973, p. 798).

Proposition 2. *When $\alpha = 1$ in* (B.2)*, the t-ratio $(\widehat{\delta} - \delta)/\widehat{\sigma}$ converges weakly to a nondegenerate limiting distribution only if S_g belongs to either the domain of attraction of a Cauchy law or a translate of Cauchy law. Hence, no confidence set constructed using quantiles of the asymptotic distribution of the t-ratio can achieve uniform size control over* $\mathbf{P}(0)$.

Nonetheless, we show a next best result holds true: the proposed inference can control size uniformly over the set $\mathbf{P}(\varepsilon)$ if $\varepsilon > 0$. Recall that if the weight vectors $(w_1^j, ..., w_G^j)$ in the bootstrap are instead generated following sampling b out of G unit without placement, then the proposed procedure becomes subsampling. Denote the empirical CDFs for the complete subsampling

$$L_G(x, P) = \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ \frac{\widehat{\delta}_{b,j} - \delta}{\widehat{\sigma}_{b,j}} \leqslant x \right\}, \quad \widehat{L}_G(x) = \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ \frac{\widehat{\delta}_{b,j} - \widehat{\delta}}{\widehat{\sigma}_{b,j}} \leqslant x \right\},$$

where $B_G = \binom{G}{b}$. Further, let the a-th quantile of $\widehat{L}_G(\cdot)$ be denoted by $\widehat{L}_G^{-1}(a)$.

Theorem 6 (Uniformity)**.** *For any $\varepsilon \in (0, 1]$, the confidence sets constructed based on the proposed subsampling procedure achieves asymptotically uniform size control over $\mathbf{P}(\varepsilon)$. Explicitly, for any nonnegative a_1 and a_2 such that $0 \leqslant a_1 + a_2 < 1$, we have*

$$\lim_{G \to \infty} \inf_{P \in \mathbf{P}} P \left(\widehat{L}_G^{-1}(a_1) \leqslant \frac{\widehat{\delta} - \delta}{\widehat{\sigma}} \leqslant \widehat{L}_G^{-1}(1 - a_2) \right) = 1 - a_1 - a_2.$$

Furthermore, if $b^2/G = o(1)$, the same conclusion holds for CS bootstrap in place of subsampling.

A proof can be found in Appendix C.6. The proof utilizes the general results in Romano and Shaikh (2012) under high-level conditions together with our Lemma 2 in Appendix C.6. This new lemma establishes a novel convergence in distribution result for row-wise i.i.d.

triangular arrays. Specifically, we consider the sequence of indices $\alpha_G \to \alpha_0 \in [1+\varepsilon, 2]$ as $G \to \infty$, covering the cases with both normal ($\alpha_0 = 2$) and non-Gaussian ($\alpha_0 < 2$) limiting distributions. Recall that the t-test is not uniformly valid over the set of all DGPs with finite second moments, while it controls size uniformly over the set of all DGPs with finite $2+\epsilon$ moments for any $\epsilon > 0$ (see e.g. Romano 2004). Our result with $\mathbf{P}(\varepsilon)$ for all $\varepsilon > 0$ is analogous to this classic result, although it extends the scope of uniformity to a much larger class of DGPs with potentially infinite second moments and non-Gaussian limiting distributions.

Furthermore, we note that it is likely that the additional condition $b^2/G = o(1)$ imposed for the CS bootstrap can be removed by appealing to, e.g. the generic results of Andrews, Cheng, and Guggenberger (2020). The present argument, for its simplicity, relies on exploiting the asymptotic equivalence between subsampling and the m-out-of-n bootstrap under the regime $m^2/n = o(1)$, together with the high-level uniformity result for subsampling in Romano and Shaikh (2012), and is therefore not optimally tailored to the CS bootstrap setting. We leave a refinement of the proof along these lines for future research.

Finally, it is noteworthy that our uniform size control property exhibits resemblances to certain instances in the existing literature. An example is the unit root model presented in Example 1 of Andrews et al. (2020), where uniform size control persists across DGPs leading to either normal or non-Gaussian limiting distributions. In that example, Andrews et al. (2020) demonstrate the continuity of their limiting distribution in a local parameter h throughout its support, akin to the role served by our nuisance parameters (α, p) in our asymptotic theory. Notably, while infinite variance poses no hindrance in Andrews et al. (2020), its presence significantly complicates the analytical framework within our study. To the best of our knowledge, Theorem 6 stands as the first theoretical result addressing the uniformity property of subsampling or bootstrap for statistical models that may exhibit potentially infinite variance.

C Mathematical Proofs

This section collects all the mathematical proofs. The order in which the proofs appear differs from that of the corresponding statements in the main text for the following reasons. The proof of Theorem 2 relies on Lemma 1, and hence we present the proof of Lemma 1 before that of Theorem 2. Similarly, the proof of Theorem 1 depends on both Lemma 1 and Theorem 2, so we present those results before turning to Theorem 1. Proofs for all remaining theorems are given in the order of appearance of their corresponding statements, namely, Theorem 5 (failure of the wild bootstrap) and Theorem 6 (uniformity). Some additional technical lemmas are relegated to Appendix D.

C.1 Proof of Lemma 1

Proof of Lemma 1. We establish the result for both the CS bootstrap and subsampling. 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 CS bootstrap.

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 $\sum_{g=1}^{G} S_g$:

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

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

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

as $G \to \infty$, where

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

$Z_k = (E_1 + ... + E_k)^{-1/\alpha}$ 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 (C.2). By Theorems 1 and 1′ in LePage et al. (1981),

$$\begin{aligned}&\left(\frac{1}{A_G}\sum_{g=1}^{G} S_g, \frac{1}{A_G^2}\sum_{g=1}^{G} S_g^2\right)\\ &\xrightarrow{d}(S,V) := \left(\sum_{k=1}^{\infty}\{\epsilon_k Z_k - (2p-1)\mathbb{E}[Z_k\mathbb{1}(Z_k < 1)]\}, \sum_{k=1}^{\infty} Z_k^2\right) = O_p(1)\end{aligned} \tag{C.3}$$

holds for $A_G = G^{1/\alpha}L_1(G)$, where Z_k, ϵ_k, and p are defined below Equation (C.2), and $L_1(\cdot)$ is slowly varying at ∞; and

$$\frac{1}{(A'_G)^2}\sum_{g=1}^{G}(X'_g X_g)^2 \xrightarrow{d} \sum_{k=1}^{\infty} \widetilde{Z}_k^2 = O_p(1) \tag{C.4}$$

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 (C.3) implies the consistency

$$\|\widehat{\theta} - \theta\| = \left\|\left(\sum_{g=1}^{G} X'_g X_g\right)^{-1} \sum_{g=1}^{G} S_g\right\| = O_p(L_1(G)G^{-(1-1/\alpha)}) = o_p(1) \tag{C.5}$$

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,

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

holds, where the second inequality follows from the Cauchy-Schwarz inequality and the stochastic orders are due to Equations (C.3), (C.4), and (C.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 the CS bootstrap and subsampling. Define the regular bootstrapped estimator

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

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

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

where

$$\left\|\frac{1}{G}\sum_{g=1}^{G} X_g X_g - \frac{1}{b}\sum_{g=1}^{G} w_g^j X_g' X_g\right\| = o_p(1)$$

follows from an application of the main theorem in Csörgo (1992) under his Condition (1.5). This implies $\widehat{\theta}_{b,j} = \breve{\theta}_{b,j}(1 + o_p(1))$. Therefore, in the bootstrap/subsampling process, $\breve{\theta}_{b,j}$ can be replaced by $\widehat{\theta}_{b,j}$ without changing the asymptotic behavior. Thus, it suffices to establish the validity of the CS bootstrap procedure based on the conventional m-out-of-n 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 4.1 in Arcones and Giné (1991), we have

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

as $M \to \infty$ and $G \to \infty$. Step 1, the triangle inequality, and the continuity of $J^*(\cdot)$ then imply

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

This concludes the proof for the bootstrap.

The proof for subsampling proceeds in the same way above with with Csörgo (1992) no longer needed and Theorem 4.1 in Arcones and Giné (1991) replaced by Theorem 11.3.1 in Politis, Romano, and Wolf (1999). □

C.2 Proof of Theorem 2

Proof of Theorem 2. We establish the result for both the CS bootstrap and subsampling. The case of $\alpha < 2$ follows directly from Lemma 1. For $\alpha = 2$, the proof is similar to the proof of Lemma 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

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

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

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

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

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

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 Lemma 1 still go through, and thus for the self-normalized sums defined in Equation (C.1), it holds that $SN_2(\theta) = SN_1(\theta) + o_p(1)$. Finally, for the validity of the CS bootstrap, we now invoke Theorem 4.1 in Arcones and Giné (1991) for the bootstrap case and Theorem 2.2.1 in Politis et al. (1999) for subsampling. Note that the limiting distribution is normal and hence continuous. □

C.3 Proof of Theorem 1

Proof of Theorem 1. The if part of the statement follows from the proof of Theorem 2. 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 (C.1) satisfy $SN_2(\theta) = SN_1(\theta) + o_p(1)$, as shown in the proofs for Lemma 1 and Theorem 2. □

C.4 Proof of Theorem 3

Proof of Theorem 3. Since we can condition on the realization of b and take random subsets of clusters, it suffices to show that the data-driven choice b as in Algorithm 1 satisfies that as $G \to \infty$, it holds with probability approaching one that

$$\widehat{b} \to \infty \text{ and } \widehat{b}/G \to 0. \tag{C.6}$$

It suffices to prove the statement for the univariate self-normalized sum

$$T_G = \frac{\sum_g S_g}{\sqrt{\sum_g S_g^2}},$$

as the general case follows under Assumption 1. In particular, under Assumption 1, the design matrix is consistent for its limit and, as shown in the proof of Lemma 1, replacing S_g with $\widehat{S}_g$ does not affect the asymptotic distribution.

Note that because the function ϕ used in the construction of b_ℓ is sublinear in Algorithm 1, we have $\widehat{b}/G \leqslant \lceil q \cdot \phi(G) \rceil / G \to 0$. Then it remains to show $\widehat{b} \to \infty$ with probability $1 - o(1)$.

To this end, we follow the same argument as in the first half of the proof of Theorem 1 in Bickel and Sakov (2008), from the beginning up to their Equation (22), by verifying Conditions (A.1)–(A.4) in that paper. Also notice that excluding the case $b \sim G$, as in our

Algorithm 1, will not affect these parts of the argument.

Condition (A.1) is immediate following the fact that with probability one, the mapping from data to $T_{G,b}$ is continuous. Conditions (A.2) and (A.3) are shown in the proof of our Lemma 1.

To check (A.4), for each $k = 1, ..., G$, define

$$T_k = \frac{\overline{S}_k}{\sqrt{\overline{Q}_k}} = \frac{\sum_{g=1}^k S_g}{\sqrt{\sum_{g=1}^k S_g^2}}, \qquad k \in \mathbb{N},$$

and denote by L_k the CDF of T_k when $S_1, \dots, S_k \overset{\text{i.i.d.}}{\sim} F$, where F is a non-degenerate distribution in the domain of attraction of an α-stable law for an $\alpha \in (1, 2]$. Condition (A.4) requires the mapping $k \mapsto L_k$ to be injective. We now show that $L_k \neq L_\ell$ whenever $k \neq \ell$. For any $\mathbf{s} = (s_1, \dots, s_k) \in \mathbb{R}^k$, Cauchy-Schwarz inequality implies that the support of L_k is contained in $[-\sqrt{k}, \sqrt{k}]$. Let $0 < \varepsilon < \sqrt{k}$. Because F is non-degenerate, there exist $a > 0$ such that for any $\delta \in (0, 1)$, we have $\overline{p}(\delta) := P\big(S_g \in (a, (1 + \delta)a)\big) > 0$. Consider the event

$$E_k(\delta) = \{S_1, \dots, S_k \in (a, (1 + \delta)a)\}.$$

Observe that $P(E_k(a, \delta)) = \overline{p}(\delta)^k > 0$. Furthermore, conditionally on $E_k(a, \delta)$, we have $\overline{S}_k \in (ka, k(1 + \delta)a)$ and $\overline{Q}_k \in (ka^2, k(1 + \delta)^2 a^2)$, so $T_k \in (\sqrt{k}/(1 + \delta), (1 + \delta)\sqrt{k})$. Choosing ε so that $\delta < \varepsilon/\sqrt{k}$ ensures $T_k > \sqrt{k} - \varepsilon$ on $E_k(\delta)$. Hence,

$$P\big(|T_k| > \sqrt{k} - \varepsilon\big) \geqslant \overline{p}(\delta)^k > 0.$$

Since δ (and hence ε) can be made arbitrarily small, this implies that $\operatorname{ess\,sup} |T_k| = \sqrt{k}$. Take two integers $k < \ell$ and define $A_k = \{|t| > \sqrt{k}\}$. Note that $L_k(A_k) = 0$. Also, for $\varepsilon = (\sqrt{\ell} - \sqrt{k})/2$, $L_\ell(A_k) > 0$, so $L_k \neq L_\ell$. Since the argument holds for all such k and ℓ, the map $k \mapsto L_k$ is injective, showing that Condition (A.4) is satisfied. □

C.5 Proof of Theorem 5

Proof of Theorem 5. 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 P denote the probability measure for the data and P^* denote the probability measure of Rademacher auxiliary random variables. Define

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

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 Lemma 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,

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

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

$$\mathbb{E}^* \left[\epsilon_k Z_k v_k^*\right] = 0 \text{ for all } k \text{ and}$$
$$\sum_{k=1}^{\infty} \mathbb{E}^* \left[(\epsilon_k Z_k v_k^*)^2\right] = \sum_{k=1}^{\infty} Z_k^2 < \infty$$

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 $k \to \infty$ 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 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 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 P and the right-hand side converges to zero as $\varepsilon \to 0$, which implies via Markov's inequality that, for any $\delta > 0$,

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

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

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

P-almost surely, which follows immediately from the fact that

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

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 P-probability one. On the other hand, recall from Step 1 in the proof of Lemma 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 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 Lemma 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 P-probability.

Finally, note that $V_G^* > 0$ 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 yield 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 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 P-probability. The inconsistency then follows. □

C.6 Proof of Theorem 6

Proof of Theorem 6. Let us first introduce the following lemma.

Lemma 2 (Weak convergence of triangular arrays)**.** *For any sequence of $P_G \in \mathbf{P}(\varepsilon)$ such that $\alpha_G \to \alpha_0 \in [1 + \varepsilon, 2]$ and $p_G \to p_0 \in [0, 1]$ as $G \to \infty$, we have*

$$R_{1G} \xrightarrow{d} \mathbb{S}_{\alpha_0, p_0}.$$

Its proof is presented in the end of this section.

Now, to show the statement of Theorem 6, we shall derive the asymptotic distribution for the following self-normalized sums of S_g:

$$R_{1G} := \frac{\sum_{g=1}^{G} S_g}{\sqrt{\sum_{g=1}^{G} S_g^2}} \quad \text{and} \quad R_{2G} := \frac{\widehat{\delta} - \delta}{\widehat{\sigma}} = \frac{\sum_{g=1}^{G} S_g}{\sqrt{\sum_{g=1}^{G} \widehat{S}_g^2}}. \tag{C.7}$$

Following Eq (1.3) in Logan et al. (1973), we obtain

$$R_{2G} = R_{1G}\left(\frac{G}{G - R_{1G}^2}\right)^{1/2}.$$

Thus, by Lemma 2, the limiting distribution of R_{2G} coincides with the one of R_{1G}.

The proof follows a similar structure to the one for Theorem 3.1 in Romano and Shaikh (2012). We will apply our Lemma 3 in Appendix D with

$$R_G = \frac{\widehat{\delta} - \delta}{\widehat{\sigma}} \quad \text{and} \quad \widehat{R}_b = \frac{\widehat{\delta}_{b,j} - \widehat{\delta}}{\widehat{\sigma}_{b,j}}.$$

First, we verify

$$\sup_{P \in \mathbf{P}} \sup_{x \in \mathbb{R}} |J_b(x, P) - J_G(x, P)| \to 0 \tag{C.8}$$

as $b, G \to \infty$ with $b/G = o(1)$. By way of contradiction, assume that it fails. Then, there exists a subsequence G_l and some $(\alpha, p) \in [1 + \varepsilon, 2] \times [0, 1]$ such that either

$$\sup_{x \in \mathbb{R}} |J_{b_{G_l}}(x, P_{G_l}) - F_{\alpha,p}(x)| \nrightarrow 0 \quad \text{or} \quad \sup_{x \in \mathbb{R}} |J_{G_l}(x, P_{G_l}) - F_{\alpha,p}(x)| \nrightarrow 0.$$

Recall that $\mathbb{S}_{\alpha,p} \sim F_{\alpha,p}$ has a continuous distribution (almost everywhere). Yet, either of these would violate Lemma 2. Thus Condition (C.8) must hold.

We will next verify the condition that

$$\sup_{P \in \mathbf{P}} P\left(\sup_{x \in \mathbb{R}} \left|\widehat{L}_G(x) - L_G(x, P)\right| > \varepsilon'\right) = o(1)$$

for all $\varepsilon' > 0$. Consider any sequence $\{P_G \in \mathbf{P} : G \geqslant 1\}$. For any $\eta > 0$, we have

$$\begin{aligned}
&\sup_{x \in \mathbb{R}} \{\widehat{L}_G(x) - L_G(x, P_G)\} \\
\leqslant &\sup_{x \in \mathbb{R}} \{\widehat{L}_G(x) - L_G(x + \eta, P_G)\} + \sup_{x \in \mathbb{R}} \{L_G(x + \eta, P_G) - L_G(x, P_G)\} \\
\leqslant &\sup_{x \in \mathbb{R}} \{\widehat{L}_G(x) - L_G(x + \eta, P_G)\} + \sup_{x \in \mathbb{R}} \{L_G(x + \eta, P_G) - L_b(x + \eta, P_G)\} \\
+ &\sup_{x \in \mathbb{R}} \{L_b(x, P_G) - L_G(x, P_G)\} + \sup_{x \in \mathbb{R}} \{L_b(x + \eta, P_G) - L_b(x, P_G)\}
\end{aligned}$$

$$= (i) + (ii) + (iii) + (iv).$$

Note that (ii) and (iii) are both $o_{P_G}(1)$ by Lemma 4.5 in Romano and Shaikh (2012). Furthermore, (iv) converges to zero as $\eta \to 0$.

Finally, we will verify $(i) = o_{P_G}(1)$ as $\eta \to 0$. By considering a subsequence, if necessary, one may assume without loss of generality that P_G is such that $\alpha_G \to \alpha$ and $p_G \to p$. The proof for this statement utilizes an argument similar to those taken in Theorem 11.3.1 in Politis et al. (1999). By its definition,

$$\begin{aligned}
\widehat{L}_G(x) =& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ \frac{\widehat{\delta}_{b,j} - \widehat{\delta}}{\widehat{\sigma}_{b,j}} \leqslant x \right\} \\
\leqslant& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ \frac{\widehat{\delta}_{b,j} - \delta}{\widehat{\sigma}_{b,j}} \leqslant x + \frac{\widehat{\delta} - \delta}{\widehat{\sigma}_{b,j}} \right\} \\
\leqslant& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ \frac{\widehat{\delta}_{b,j} - \delta}{\widehat{\sigma}_{b,j}} \leqslant x + \eta \right\} + (1 - R_G(\eta)),
\end{aligned}$$

where $R_G(\eta)$ is defined for $\eta > 0$ as

$$\begin{aligned}
R_G(\eta) =& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ \frac{\widehat{\delta} - \delta}{\widehat{\sigma}_{b,j}} \leqslant \eta \right\} \\
=& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ (b/A_b)\widehat{\sigma}_{b,j} \geqslant (b/A_b)(\widehat{\delta} - \delta)/\eta \right\},
\end{aligned}$$

$A_b = b^{1/\alpha} L(b)$ for some slow varying L at infinity. As $A_G/A_b \to 0$, for any $\varepsilon'' > 0$, it holds that $(b/A_b)(\widehat{\delta} - \delta) \leqslant \varepsilon''$ with probability approaching one along P_G. This is because $\widehat{\delta}$ is the full sample estimator and thus $(G/A_G)(\widehat{\delta} - \delta) = O_{P_G}(1)$ follows from the proof of Lemma 2. As such, following the proof of Lemma 2, we have

$$R_G(\eta) \geqslant \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1} \left\{ (b/A_b)\widehat{\sigma}_{b,j} \geqslant \varepsilon''/\eta \right\} \overset{P_G}{\to} P_G(V \geqslant \varepsilon''/\eta)$$

as $G \to \infty$, where V is the stable distribution with index of stability of $\alpha/2$. By Theorem $1'$ in LePage et al. (1981) for example, V has the representation $V = \sum_{k=1}^{\infty} Z_k^2$, where

$Z_k = (E_1 + ... + E_k)^{-1/\alpha}$ for each k, $\{E_k\}_k$ are i.i.d. standard exponential random variables. As ε'' can be arbitrarily small, we have $R_G(\eta) = 1 + o_{P_G}(1)$. Thus, we have

$$\begin{aligned}\widehat{L}_G(x) \leqslant & \frac{1}{B_G}\sum_{j=1}^{B_G} \mathbb{1}\left\{\frac{\widehat{\delta}_{b,j} - \delta}{\widehat{\sigma}_{b,j}} \leqslant x + \eta\right\} + (1 - R_G(\eta)) \\ \leqslant & L_G(x + \eta, P_G) + o_{P_G}(1).\end{aligned}$$

A similar argument derives $\widehat{L}_G(x) \geqslant L_G(x + \eta, P_G) + o_{P_G}(1)$. This shows $(i) = o_{P_G}(1)$ as $\eta \to 0$, and hence concludes the proof of Theorem 6 for the case of subsampling.

Finally, under the regime of $b^2/G \to 0$, the result follows from the asymptotic equivalence between subsampling and the m-out-of-n bootstrap under the regime $m^2/n \to 0$; see Corollary 2.3.1 in Politis et al. (1999). □

Proof of Lemma 2. First, consider the case of $\alpha_0 < 2$. Denote $S_g = S_g(\alpha, p)$ to emphasize the dependence of the DGP on the index α of stability and the tail balancing parameter p. (It does not suggest that the DGP is uniquely defined by these two parameters.) For each DGP, $P_G \in \{P_G : G \geqslant 1\} \subset \mathbf{P}_1(\varepsilon)$, with indices (α_m, p_m) for an auxiliary index $m = G$, define

$$X_{mn} = \frac{\sum_{g=1}^n S_g(\alpha_m, p_m)}{\sqrt{\sum_{g=1}^n S_g^2(\alpha_m, p_m)}}$$

for each $n \geqslant 1$. Since (α_m, p_m) is fixed over n for each m, we can apply Theorem 1′ in LePage et al. (1981) to obtain that, for each m as $n \to \infty$, there exists some positive sequence $A_{mn} \to \infty$ such that

$$\begin{aligned}&\left(\frac{1}{A_{mn}}\sum_{g=1}^n S_g(\alpha_m, p_m), \frac{1}{A_{mn}^2}\sum_{g=1}^n S_g^2(\alpha_m, p_m)\right) \\ &\xrightarrow{d} \left(\sum_{k=1}^\infty \{\epsilon_k(p_m) Z_k(\alpha_m) - (2p_m - 1)\mathbb{E}[Z_k(\alpha_m)\mathbb{1}(Z_k(\alpha_m) < 1)]\}, \sum_{k=1}^\infty Z_k^2(\alpha_m)\right) = (S_m, V_m)\end{aligned}$$

as $n \to \infty$, where $Z_k(\alpha_m) = (E_1 + ... + E_k)^{-1/\alpha_m}$ for each k, $\{E_k\}_k$ are i.i.d. standard exponential random variables, and $\{\epsilon_k(p_m)\}_k$ are i.i.d. random variables that take the value of 1 with probability p_m and -1 with probability $(1-p_m)$ and are independent of $\{Z_k(\alpha_m)\}_k$. Note that the distributions of both S_m and V_m are stable with indices of stability of α_m and $\alpha_m/2$, respectively. Furthermore, it follows from Corollary 1 in LePage et al. (1981) that

$$X_{mn} \xrightarrow{d} X_m \overset{d}{=} \frac{\sum_{k=1}^{\infty}\{\epsilon_k(p_m)Z_k(\alpha_m) - (2p_m - 1)\mathbb{E}[Z_k(\alpha_m)\mathbb{1}(Z_k(\alpha_m) < 1)]\}}{\sqrt{\sum_{k=1}^{\infty} Z_k^2(\alpha_m)}}.$$

Let the limiting distribution on the right-hand side be denoted by $\mathbb{S}_{\alpha_m,p_m}$. Also, note that $(\alpha_m, p_m) \to (\alpha_0, p_0)$ by our construction, and thus,

$$X_m \xrightarrow{d} X \sim \mathbb{S}_{\alpha_0,p_0}$$

follows from the convergence of the sequence of the characteristic functions of the stable distributions S_m and V_m, as these characteristic functions are continuous in (α, p) over $(1, 2) \times [0, 1]$ (cf. Remark 4 on page 7 in Samorodnitsky and Taqqu, 1994) and V_m is positive with probability one for all $\alpha \in (1, 2)$.

Next, by invoking the Skorohod's representation theorem (as $\mathbb{R}$ is a separable metric space), there exist versions of X_{mn} and X_m such that $X_{mn} \xrightarrow{a.s.} X_m$ for each m and as $n \to \infty$, and $X_m \xrightarrow{a.s.} X$ as $m \to \infty$. Now, for such X_{mn}, define $Y_n = X_{nn}$. By construction, we have $Y_n \overset{d}{=} R_{1n}$ for all $n \geqslant 1$. Also, it follows from the almost sure converges that

$$\lim_{M\to\infty} \limsup_{n\to\infty} P(|X_{mn} - Y_n| \geqslant \varepsilon) = 0$$

for all $\varepsilon > 0$. Applying Lemma 4 in Appendix D, we have $Y_n \xrightarrow{d} X$ as $n \to \infty$. Thus, we conclude $R_{1n} \xrightarrow{d} X$.

Now, consider the case of $\alpha_0 = 2$. We only need to consider the case where we have $\alpha_G < 2$ for at least one G, as, otherwise, $\alpha_G = 2$ for all G and

$$R_{1G} \xrightarrow{d} \mathcal{N}(0, 1)$$

follows immediately from the Lindeberg-Feller CLT. Now, for those $\alpha_m < 2$, construct X_{mn}

as in the previous case. By Corollary in LePage et al. (1981), we have

$$X_{mn} = \frac{\sum_{g=1}^{n} S_g(\alpha_m, p_m)}{\sqrt{\sum_{g=1}^{n} S_g^2(\alpha_m, p_m)}} \xrightarrow{d} X_m \sim \mathbb{S}_{\alpha_m, p_m}.$$

By Assertion (vi) in Section 5 and Equation (5.13) in Logan et al. (1973), the density $f_{\alpha_m,p_m}(\cdot)$ of $\mathbb{S}_{\alpha_m,p_m}$ exists and is bounded everywhere except on a set with measure zero, and, as $\alpha_m \to 2$, $f_{\alpha_m,p_m} \to \varphi$, the standard normal density, on the real line. Thus, by the bounded convergence theorem, the CDF $F_{\alpha_m,p_m}(x)$ of $\mathbb{S}_{\alpha_m,p_m}$ converges to the standard normal distribution function $\Phi(x)$ for all $x \in \mathbb{R}$, i.e. $X_m \xrightarrow{d} X \sim \mathcal{N}(0,1)$. Using the same construction of Y_n as above, we conclude $R_{1n} \xrightarrow{d} \mathcal{N}(0,1)$ by Lemma 4 in Appendix D. □

D Auxiliary Lemmas from the Literature

The following lemma restates Theorems 2.1 and 2.2 as well as Remark 2.1 in Romano and Shaikh (2012) for convenience of reference.

Lemma 3 (High-level uniformity)**.** *For subsampling and under the setup in Section B,*

$$\lim_{G\to\infty} \sup_{P\in\mathbf{P}} \sup_{x\in\mathbb{R}} |J_b(x,P) - J_G(x,P)| = 0,$$

implies

$$\liminf_{G\to\infty} \inf_{P\in\mathbf{P}} P\left(L_G^{-1}(a_1,P) \leqslant R_G \leqslant L_G^{-1}(1-a_2,P)\right) \geqslant 1 - a_1 - a_2$$

for any nonnegative a_1 and a_2 such that $0 \leqslant a_1 + a_2 < 1$. In addition, if $J_G(x,P)$ tends in distribution to a limiting distribution $J(x,P)$ that is continuous, then

$$\lim_{G\to\infty} \inf_{P\in\mathbf{P}} P\left(L_G^{-1}(a_1,P) \leqslant R_G \leqslant L_G^{-1}(1-a_2,P)\right) = 1 - a_1 - a_2.$$

Finally, if

$$\sup_{P\in\mathbf{P}} P\left(\sup_{x\in\mathbb{R}} \left|\widehat{L}_G(x) - L_G(x,P)\right| > \varepsilon\right) = o(1)$$

for all $\varepsilon > 0$, then

$$\lim_{G\to\infty} \inf_{P\in\mathbf{P}} P\left(\widehat{L}_G^{-1}(a_1) \leqslant R_G \leqslant \widehat{L}_G^{-1}(1-a_2)\right) = 1 - a_1 - a_2.$$

The next result is identical to Theorem 3.5 in Resnick (2007).

Lemma 4 (Second converging together theorem)**.** *Suppose that $\{X_{mn}, X_m, X, Y_n : n \geqslant 1, m \geqslant 1\}$ are random elements of the metric space $(\mathbb{S}, \mathcal{S})$ with a metric $d(\cdot,\cdot)$ that are defined on a common domain. Assume that for each m, as $n \to \infty$, $X_{mn} \rightsquigarrow X_m$, and as $m \to \infty$, $X_m \rightsquigarrow X$, Further suppose that for all $\varepsilon > 0$,*

$$\lim_{m\to\infty} \limsup_{n\to\infty} P(d(X_{mn}, Y_n) \geqslant \varepsilon) = 0.$$

Then, as $n \to \infty$, we have $Y_n \rightsquigarrow X$, where $\rightsquigarrow$ denotes weak convergence.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023): "When should you adjust standard errors for clustering?" *Quarterly Journal of Economics*, 138, 1–35.

Akhtari, M., D. Moreira, and L. Trucco (2022): "Political turnover, bureaucratic turnover, and the quality of public services," *American Economic Review*, 112, 442–493.

Andrews, D. W., X. Cheng, and P. Guggenberger (2020): "Generic results for establishing the asymptotic size of confidence sets and tests," *Journal of Econometrics*, 218, 496–531.

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.

——— (1991): "Additions and correction to "The bootstrap of the mean with arbitrary bootstrap sample"," in *Annales de l'IHP Probabilités et statistiques*, vol. 27, 583–595.

Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

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

Bai, Y., J. Liu, A. M. Shaikh, and M. Tabord-Meehan (2022): "Inference in Cluster Randomized Trials with Matched Pairs," *arXiv preprint arXiv:2211.14903*.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Bickel, P. J. and A. Sakov (2008): "On the choice of m in the m out of n bootstrap and confidence bounds for extrema," *Statistica Sinica*, 967–985.

Bugni, F., I. Canay, A. Shaikh, and M. Tabord-Meehan (2024): "Inference for cluster randomized experiments with non-ignorable cluster sizes," *Journal of Political Economy Microeconomics*, Forthcoming.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *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.

——— (2025): "Inference for regression with clustered and spatially correlated data," Online; slides, available at https://cameron.econ.ucdavis.edu/research/papers.html.

Canay, I. A., A. Santos, and A. M. Shaikh (2021): "The wild bootstrap with a "small" number of "large" clusters," *Review of Economics and Statistics*, 103, 346–363.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017): "Asymptotic behavior of at-test robust to cluster heterogeneity," *Review of Economics and Statistics*, 99, 698–709.

Cavaliere, G., I. Georgiev, and R. A. Taylor (2013): "Wild bootstrap of the sample mean in the infinite variance case," *Econometric Reviews*, 32, 204–219.

Cavaliere, G., T. Mikosch, A. Rahbek, and F. Vilandt (2024): "Tail behavior of ACD models and consequences for likelihood-based estimation," *Journal of Econometrics*, 238, 105613.

Chernozhukov, V., I. Fernández-Val, and T. Kaji (2016): "Extremal quantile regression: An overview," *arXiv preprint arXiv:1612.06850.*

——— (2017): "Extremal quantile regression," *Handbook of Quantile Regression*, 333–362.

Cornea-Madeira, A. and R. Davidson (2015): "A parametric bootstrap for heavy-tailed distributions," *Econometric Theory*, 31, 449–470.

Csörgo, S. (1992): "On the law of large numbers for the bootstrap mean," *Statistics & probability letters*, 14, 1–7.

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.

Eeckhout, J. (2004): "Gibrat's law for (all) cities," *American Economic Review*, 94, 1429–1451.

Embrechts, P. and C. M. Goldie (1980): "On closure and factorization properties of subexponential and related distributions," *Journal of the Australian Mathematical Society*, 29, 243–256.

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

Geluk, J. L. and L. de Haan (2000): "Stable probability distributions and their domains of attraction: a direct approach," *Probability and Mathematical Statistics-Wroclaw Univeristy*, 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. (2022a): *Econometrics*, Princeton University Press.

Hansen, B. E. (2022b): "Jackknife standard errors for clustered regression," *Working Paper.*

——— (2022c): "Jackknife standard errors for clustered regression," *University of Wisconsin.*

Hansen, B. E. and S. Lee (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268–290.

Hansen, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141, 597–620.

Hersch, J. (1998): "Compensating differentials for gender-specific job injury risks," *American Economic Review*, 88, 598–607.

Hirano, K. (1998): "Keisuke Hirano: Haiku," Accessed: 18 September 2025.

Hirano, K. and J. R. Porter (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 80, 1769–1790.

Ibe, O. (2013): *Markov Processes for Stochastic Modeling*, Newnes.

Ibragimov, M., R. Ibragimov, and J. Walden (2015): *Heavy-tailed distributions and robustness in economics and finance*, vol. 214, Springer.

Ibragimov, R. and U. K. Müller (2010): "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28, 453–468.

——— (2016): "Inference with few heterogeneous clusters," *Review of Economics and Statistics*, 98, 83–96.

Ioannides, Y. and S. Skouras (2013): "US city size distribution: Robustly Pareto, but only in the tail," *Journal of Urban Economics*, 73, 18–29.

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.

Liang, K.-Y. and S. L. Zeger (1986): “Longitudinal data analysis using generalized linear models,” *Biometrika*, 73, 13–22.

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 (2022): “Fast and reliable jackknife and bootstrap methods for cluster-robust inference,” *Working Paper.*

——— (2023): “Cluster-robust inference: A guide to empirical practice,” *Journal of Econometrics*, 232, 272–299.

Menzel, K. (2021): “Bootstrap with cluster-dependence in two or more dimensions,” *Econometrica*, 89, 2143–2188.

Mikosch, T. (1999): *Regular variation, subexponentiality and their applications in probability theory*, vol. 99, Eindhoven University of Technology Eindhoven, The Netherlands.

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. (2004): “On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems,” *Scandinavian Journal of Statistics*, 31, 567–584.

Romano, J. P. and A. M. Shaikh (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40, 2798–2822.

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 (2023): "Diagnostic Testing of Finite Moment Conditions for the Consistency and Root-N Asymptotic Normality of the GMM and M Estimators," *Journal of Business & Economic Statistics*, 41, 339–348.

White, H. (1984): *Asymptotic Theory for Econometricians*, Academic Press.