Genuinely Robust Inference for Clustered Data[*]

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

Abstract

Conventional methods for cluster-robust inference are inconsistent when clusters of unignorably large size are present. We formalize this issue by deriving a necessary and sufficient condition for consistency, a condition frequently violated in empirical studies. Specifically, 77% of empirical research articles published in *American Economic Review* and *Econometrica* during 2020–2021 do not satisfy this condition. To address this limitation, we propose two alternative approaches: (i) *score subsampling* and (ii) *size-adjusted reweighting*. Both methods ensure uniform size control across broad classes of data-generating processes where conventional methods fail. The first approach (i) has the advantage of ensuring robustness while retaining the original estimator. The second approach (ii) modifies the estimator but is readily implementable by practitioners using statistical software such as Stata and remains uniformly valid even when the cluster size distribution follows Zipf's law. Extensive simulation studies support our findings, demonstrating the reliability and effectiveness of the proposed approaches.

Keywords: cluster-robust inference, score subsampling, reweighting, unignorably large cluster, domain of attraction, stable distribution

JEL Code: C12, C18, C46

[*] This paper supersedes the manuscripts previously circulated under the titles "On the Inconsistency of Cluster-Robust Inference and How Subsampling Can Fix It" and "Non-Robustness of the Cluster-Robust Inference: with a Proposal of a New Robust Method." We benefited from comments and discussions with A. Colin Cameron, Ivan Canay, Bruce Hansen, Seojeong (Jay) Lee, James MacKinnon, Francesca Molinari, Ulrich Müller, Jack Porter, Frank Schorfheide, Xiaoxia Shi, Kevin Song, Max Tabord-Meehan, and seminar participants at numerous institutions and conferences. We would like to thank Hong Xu and Siyuan Xu for 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 (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 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 31 published articles.[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 $\max_g N_g^2/N \geqslant 1$ (respectively, $\geqslant 10$, $\geqslant 100$,

[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 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–	$\geqslant$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: Number of articles with $\max_g N_g^2/N$ falling in each of the bins $[0, 0.1)$, $[0.1, 1)$, $[1, 10)$, $[10, 100)$, $[100, 1000)$ and $[1000, \infty)$ in the logarithmic scale. The articles are drawn from those papers published in *American Economic Review* and *Econometrica* during the period of 2020-2021. We focus on those papers that report CR standard errors for regression and IV regression estimates with publicly available data sets for replication. For each paper running more than one regression, we take the largest $\max_g N_g^2/N$ among the regressions.

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-normal 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 3.

Non-normal 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.

To address this issue, we propose two reliable methods for clustered data: (i) subsampling-

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

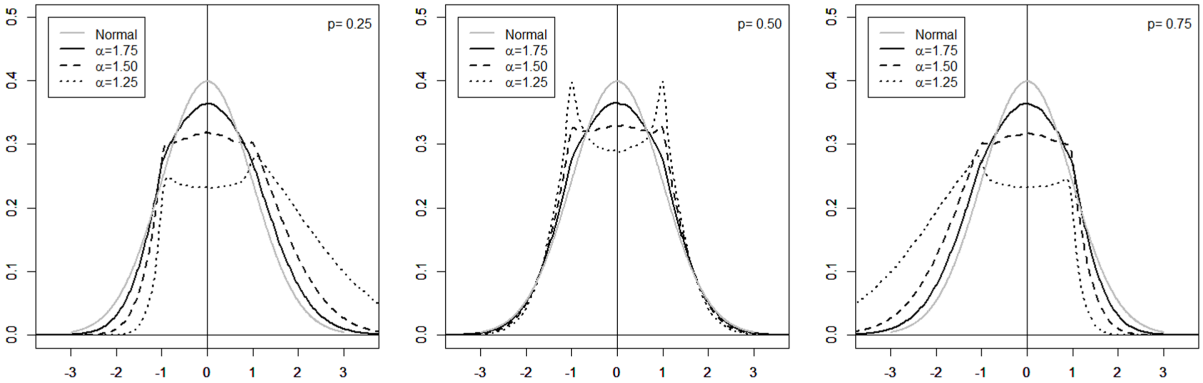

Figure 1: Illustration of non-normal limiting distributions in the presence of unignorably large clusters. Details about the different shapes indexed by α and p are found in Section 3.

	0. Conventional	1. Subsampling	2. SACR
OLS in Estimand	◎	◎	× Estimand And Estimate Change
Allows for Heavy-Tailed Cluster-Size Distributions	× Two+ Moments Are Required	◎	◎
Compatible with Zipf's Law Cluster-Size Distributions	× Two+ Moments Are Required	× One+ Moments Are Required	◎

Table 2: Advantages and disadvantages of alternative approaches to cluster-robust inference. While the existing literature provides the 0. Conventional approach, we propose the 1. Subsampling and 2. SACR approaches in this paper.

based inference and (ii) size-adjusted cluster-robust (SACR) estimation. The former (i) has the advantage of retaining the original OLS estimate while the latter (ii) is both easy to implement and well-suited for cluster size distributions that follow Zipf's law. Both methods provide valid critical values adaptively across all limiting distributions depicted in Figure 1. The advantages and disadvantages of the alternative methods are summarized in Table 2.

We demonstrate that the proposed inference procedures adapt seamlessly to both normal and non-normal limiting distributions. This adaptability is established through uniform size control across a wide range of models, including data-generating processes with ignorable and notably large clusters. To ensure this uniform validity, we introduce a novel convergence-in-distribution result for row-wise i.i.d. triangular arrays with heavy tails and converging tail exponents. This represents the first theoretical result on the uniformity of subsampling for models with potentially infinite variance. Simulation studies further confirm the reliability of our methods.

Related Literature: The literature of cluster robust inference has a long history dating back to White (1984), Liang and Zeger (1986), and Arellano (1987). For a thorough review of the literature, we refer the readers to Cameron and Miller (2015) and MacKinnon, Nielsen, and Webb (2023). The sampling frameworks in which cluster sizes are treated as a random variable have been recently investigated by Bugni et al. (2024), Cavaliere, Mikosch, Rahbek, and Vilandt (2024), and Bai, Liu, Shaikh, and Tabord-Meehan (2022). We consider a model-based perspective with an increasing number of clusters and unrestricted intra-cluster dependence, as the vast majority of the papers did in this literature. An alternative framework is a fixed number of clusters with growing cluster sizes and manage to derive asymptotic normality under some extra assumptions on weak intra-cluster cluster dependence following Canay, Santos, and Shaikh (2021), as well as design-based asymptotics[3] under some stronger treatment assignment rules, such as randomized experiments, considered by Abadie, Athey, Imbens, and Wooldridge (2023).

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

[3] See Reichardt and Gollob (1999) for an in-depth philosophical discussion on the model-based versus design-based perspectives.

Our key distributional approximation results for the self-normalized sums are due to Logan, Mallows, Rice, and Shepp (1973), LePage, Woodroofe, and Zinn (1981), and Giné, Götze, and Mason (1997). For theoretical details of the underlying foundations of probability and statistics for heavy-tailed distributions, we refer the reader to Resnick (1987, 2007). Our uniformity result relies on the general uniformity theory for subsampling studied in Romano and Shaikh (2012). For the failure of empirical bootstrap for means of random variables with infinite variances, see, e.g., Athreya (1987), Arcones and Giné (1989), and Knight (1989). Our inference procedure relies on the theory of resampling method developed in Politis and Romano (1994) and Romano and Wolf (1999). Also, see Politis, Romano, and Wolf (1999) for a comprehensive treatment.

Finally, it is worth connecting our paper to the literature on economic geography and urban economics, where researchers often define clusters based on geographical regions such as states, counties, and cities. A well-established result in this literature is that city sizes follow Zipf's law (e.g., Gabaix, 1999), which implies a unit Pareto exponent. This characteristic, in turn, leads to the nonexistence of the first moment and hence non-normal limiting distributions of the least squares estimator. These statistical properties pose significant challenges for conventional inference methods. Beyond the empirical evidence we present below, this literature provides a theoretical economic foundation for the issues explored in this paper. Since our SACR method remains uniformly valid in this setting, we recommend its application in studies involving city-level clustering.

Organization Section 2 sets up the model and inference problem. Section 3 presents the fragility of the conventional methods based on formal theory and empirical examples. Section 4 introduces the subsampling method, and Section 5 introduces the SACR approach.

2 The Model

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

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

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 $\delta = r'\theta$, 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

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

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

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

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., $\mathbb{E}[N_g] = c \in (0, \infty)$, and the design matrix satisfies

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

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 less restrictive than requiring the central limit theorem to hold and even includes scenarios where the central limit theorem does not apply. We provide some discussions about Assumption 1. First, this high-level assumption accommodates a broad class of both standard and non-standard cases considered in econometrics. A low-level sufficient condition is that there exists non-trivial within-cluster dependence and that N_g follows a power law. In particular, the power law property has been established and documented for various types of clusters, such as cities and firms. See Gabaix (2009, 2016) for comprehensive reviews. We illustrate the limitations of existing methods under these conditions in Section 3.1.

Second, the case of $\alpha = 2$ encompasses the conventional assumption under which $r'(\widehat{\theta}-\theta)$ enjoys the standard convergence rate of $\sqrt{G}$ through CLTs. In this case, the limiting α-stable distribution must be normal (cf. Geluk and de Haan, 2000, Theorem 2). Furthermore, it also covers some non-standard cases with a normal limiting distribution but without a finite variance, e.g., a Pareto random variable with the shape parameter (Pareto exponent) of 2.

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

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

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 of $\alpha < 2$, the index α of stability coincides with the Pareto exponent[4] β in the sense that $\alpha = \min\{\beta, 2\}$. Thus, the case of $\alpha < 2$ implies infinite variance of the score. See Theorem 5 in Appendix A.1 for more precise details. In this case, *unignorably large* clusters are literally unignorable because the sample sum of the (scaled) scores becomes asymptotically proportional to the (scaled) score of the largest cluster - see Remark 5 in Appendix A.2 for more discussions. Hence, the asymptotic distribution cannot be normal.

As discussed in the introduction, the literature on urban economics and economic geography has established theoretical results indicating that the sizes cities follow Zipf's law (e.g., Gabaix, 1999). In particular, this implies $\alpha = 1 < 2$ when data are non-trivially clustered by cities, leading to a non-normal limit distribution.

Finally, the i.i.d. requirement in Assumption 1 is standard in this literature, (cf. Bugni et al. 2024; Cavaliere et al. 2024; Bai et al. 2022). It is mild because 1) the conditional distributions of S_g and $X_g'X_g$ given $N_g = n_g$ can be heterogeneous across n_g; and 2) the distributions of individuals within each cluster can be non-identical. In addition, S_g and X_g can be arbitrarily correlated with the cluster size N_g so long as the exogeneity condition for the regression is respected.

To simplify the writings, we focus on the case where $v'S_g$ and $u_1'X_g'X_gu_2$ share the common index α of stability. This simple setting is rationalized if the tail shape of their distributions are driven by the tail shape of the distribution of cluster sizes N_g - see Section 3.1 for an illustrating example. With this said, we emphasize that this setting is not essential and can

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

be relaxed only at the cost of more cumbersome writing.

3 Fragility of the Conventional CR Methods

In this section, we argue that the conventional methods of CR inference work if and only if $\alpha = 2$. In other words, they are doomed to fail if $\alpha < 2$. We start with some heuristic discussions in Section 3.1 and provide formal theories in Section 3.2. We also discuss how often researchers encounter cases with $\alpha < 2$ in empirical economic studies.

3.1 Some Heuristic Discussions

The intuition behind the fragility of the conventional CR method is straightforward. When $\alpha < 2$, N_g does not have a finite variance. If the intra-cluster dependence is non-trivial, the infinite variance of N_g is inherited by the score S_g, causing the CLT for the OLS (and other) estimators to fail. We believe that $\alpha < 2$ is plausible in numerous studies and showcase some recent studies published in *Econometrica* and *American Economic Review* in Section 3.2.

To illustrate this argument, let us consider the sample average

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

which is the special case of the OLS (2.1) with $X_{gi} = 1$. The true parameter value becomes the mean $\theta = \mathbb{E}[Y_{gi}]$. Without loss of generality, normalize the location to $\theta = 0$. Furthermore, let us consider the extreme case with perfect intra-cluster dependence, i.e., $Y_{gi} \equiv Y_g$ for all $i \in \{1, ..., N_g\}$ for each g. Assume N_g is independent from Y_g for simplicity. These simplifying assumptions are just for clear exposition but not essential.

In this case, we have

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

The denominator still converges to $\sqrt{\mathbb{E}\left[N_g\right]}$ provided $\alpha > 1$. For the numerator, $\mathrm{Var}[G^{-1/2}\sum_{g=1}^{G} N_g Y_g]$ is equal to $\mathrm{Var}[N_g] \cdot \mathrm{Var}[Y_g]$, which is infinite given $\alpha < 2$. In fact, Theorem 1 of Geluk and de Haan (2000) implies that, if the distribution of N_gY_g is α-stable, then the limiting distribution

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

for some sequences of constants $a_G \simeq G^{1/\alpha} \to \infty$ and $b_G \in \mathbb{R}$ 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 for $\widehat{\theta}$ fails, and the asymptotic distribution will be non-normal. We provide more discussions about this example in Appendix A.5.

In summary, the stable index α determines the convergence rate and further the asymptotic distribution. First, the tail heaviness of N_g translates to that of S_g when intra-cluster correlation is non-trivial. This results in the t-ratio of the convenctional CR method being asymptotically normal *if and only if* $\alpha = 2$ - See Theorem 1 and Proposition 1 below.

Second, when $\alpha \in (1,2)$, the convergence rate is slower than $\sqrt{G}$ and the asymptotic distribution is non-normal. However, no method based on the quantile of the asymptotic distribution of the t-ratio of the OLS estimator can uniformly control size over $\alpha \in [1,2]$ as the t-ratio fails to converge in distribution to a limit at $\alpha = 1$.

Third, an even more extreme case is when $\alpha < 1$. In such scenarios, the population problem of the OLS may not be well-defined as $\mathbb{E}[\|X_g'X_g\|]$ and $\mathbb{E}[\|S_g\|]$ fail to exist, leading to an identification failure for the OLS estimator. This echoes the conclusion in Kojevnikov and Song (2023), who show the impossibility of consistent variance estimation under the presence of a single unignorably large cluster. This scenario corresponds to an arbitrarily small α, reflecting extremely heavy tail of N_g.

Remark 1 (Bias from Trimming Large Clusters)**.** In practice, researchers may trim large clusters by randomly selecting k observations from clusters where $N_g > k$. While this approach may mitigate issues arising from non-normal limiting distributions caused by the heavy-tailed nature of N_g, it can introduce bias, thereby compromising the validity of inference. To illustrate this, consider the following:

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

Here, $\theta(k)$ represents what the trimmed estimator identifies, while $\lambda(k)$ denotes a potential bias term. This bias, $\lambda(k)$, can be nonzero if the distribution of Y_g depends on N_g. For readability, we decide to suppress details of this bias, which are available upon request.

3.2 Formal Theory

The following theorem formalizes and generalizes the discussions in the previous subsection.

Theorem 1 (Necessary and sufficient condition)**.** *Suppose that 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 found in Appendix B.3. This theorem implies that the conventional inference based on the common variance estimators, such as CR1, CR2, CR3, and jackknife, together with the normal critical values (e.g., ≈1.96 for the 97.5-th percentile) fails if $\alpha < 2$.

We can now discuss Figure 1 shown in the introductory section. Specifically, the left, middle, and right panels of Figure 1 illustrate the limiting distributions of the t-statistic under $p = 0.25$, 0.50, and 0.75, respectively, where p is the limit ratio of the tail probabilities defined as

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

for v is defined in Assumption 1. In each of these three panels, three non-normal limiting distributions corresponding to $\alpha = 1.25$, 1.50, and 1.75 are depicted with distinct line styles, along with the normal reference case ($\alpha = 2.00$). The main takeaway is that the conventional CR inference, which relies on the normal approximation, becomes increasingly size-distorted as α decreases and p, a parameter representing the limit of the tail probability ratio, deviates from 0.5.

There is certainly another class of conventional approaches, namely the cluster bootstraps. The are two main bootstrap-based CR inference methods in the literature, namely, 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), therefore, the pairs cluster bootstrap, which is essentially the empirical bootstrap performed with individual cluster-wise sums treated as the independent units, is inconsistent under Assumption 1 with $\alpha < 2$. Furthermore, we show in Theorem 6 in Appendix A.3 that the wild cluster bootstrap is also inconsistent under Assumption 1 with $\alpha < 2$. The following proposition summarizes these results.

Proposition 1 (Failure of the Cluster Bootstraps)**.** *Suppose Assumption 1 holds with $\alpha < 2$, then the pairs cluster bootstrap and the wild bootstrap methods are both inconsistent.*

Provided that the case of $\alpha < 2$ fails all these conventional methods of CR inference, our natural question now is how common it is to encounter $\alpha < 2$ in empirical studies in economics. We analyzed all the articles published in two leading journals (*American Economic Review* and *Econometrica*) between 2020 and 2021. Among them, we extracted a list of those papers that report estimation and inference results based on regressions, IV regressions, and their variants. Furthermore, we focus on those articles that use publicly available data sets for replication.

For these articles, we test the null hypothesis $H_0 : \alpha = 2$ against the alternative hypothesis $H_1 : \alpha < 2$ for the score. Such a test can be conducted via the likelihood ratio test (Sasaki and Wang, 2023) of 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.[5]

Table 3 summarizes the list of all the papers we studied. The first two columns list the journals and years of publication. The following column "All #" indicates the total number of eligible articles according to the above selection criteria. The column group under "Cluster"

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

Journal	Year of Publication	All	Cluster		
		#	#	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 3: The column "All – #" indicates the total number of eligible articles that use regressions or IV regressions with publicly available data for replication. The column "Cluster – #" indicates the number of the eligible articles that use CR inference. The column "Cluster – Test $\alpha < 2$" indicates the rate of rejecting the null hypothesis $\alpha = 2$ among those articles that use CR inference. The tests of the null hypothesis $\alpha = 2$ against the alternative hypothesis $\alpha < 2$ is implemented with the Stata command "`testout y x1 x2 ..., cluster(cid)`" for regressions and "`testout y x1 x2 ..., iv(z) cluster(cid)`" for IV regressions based on Sasaki and Wang (2023).

collects articles in which CR inference is used for at least one regression result. Under this column group, the column "#" shows the numbers of articles, and the column "Test $\alpha < 2$" shows the fractions of those articles for which the test rejects the null hypothesis for at least one regression specification. The final row displays the summary of each column.

During 2020–2021, *American Economic Review* published 30 articles meeting our selection criteria. Out of them, 21 articles report CR standard errors. We reject the null hypothesis for 16 of these 21 articles. In other words, the inference results may be misleading for 76% of those articles that employ the conventional CR method of inference.

During 2020–2021, *Econometrica* published 14 articles meeting our selection criteria. Out of them, 10 articles report CR standard errors. We reject the null hypothesis for 8 of these 9 articles. In other words, the inference results may be misleading for 80% of those articles that employ the conventional CR method of inference.

Combining two journals, we suspect potentially misleading inference results for as many as 77% of those 31 articles that employ the conventional CR method. Hence, problematic practice is prevalent even in these highly influential journals.[6]

All the above issues with the conventional CR methods motivate our proposed methods. The first method retains the OLS estimator, but uses subsampling to accommodate the non-normal limiting distribution - see Section 4. The second method, size-adjusted cluster robust (SACR) estimation, reweights the score S_g with weight N_g^{-1}, which suprresses the effects of the arbitrarily non-ignorable clusters and hence restores the CLT - see Section 5.

4 Score Subsampling as the First Reliable Solution

In light of the issue with the conventional methods of CR inference presented in the previous section, we now propose a novel method of score subsampling to approximate the limiting distribution of $(\widehat{\delta} - \delta)/\widehat{\sigma}$.

We first present the proposed method without theoretical details in Section 4.1. Its theoretical support follows in Section 4.2. Section 4.3 presents simulation studies.

4.1 The Method

Our objective is to conduct statistical inference for δ using the self-normalized t-statistic (2.3). Let the CDF J_G^* of the sampling distribution of the t-statistic be given by

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

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

$$\widehat{\delta}_{b,j} = r'\widehat{\theta}_{b,j} = \left(\frac{G}{b}\right) r' \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \sum_{g \in \mathcal{B}_j} X_g' Y_g \qquad \text{and}$$

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

$$\widehat{\sigma}_{b,j}^2 = \left(\frac{G}{b}\right)^2 r' \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\sum_{g\in\mathcal{B}_j} \widehat{S}_{g,j}\widehat{S}_{g,j}'\right)\left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} r,$$

where $\widehat{S}_{g,j} = X_g'(Y_g - X_g\widehat{\theta}_{b,j})$. Observe that the inverse factor $(\sum_{g=1}^{G} X_g'X_g)^{-1}$ is calculated based on the full sample while the linear component and its variance are computed based on the subsample $\mathcal{B}_j$. We discuss practical motivations for this feature in Remark 3 below.

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

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

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

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

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

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

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

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

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

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

Practical Implication: For the t-statistic, one can continue to use the conventional CR "standard error" $\widehat{\sigma}$.[7] However, instead of using the conventional critical values, $\Phi^{-1}(0.025) \approx$

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

-1.96 and $\Phi^{-1}(0.975) \approx 1.96$, one should use $\widehat{c}_{G,b}(0.025)$ and $\widehat{c}_{G,b}(0.975)$ obtained by the score subsampling to construct a 95% confidence interval for example.

Remark 2 (Practicality of the method)**.** The convergence rate of $\widehat{\delta} - \delta$ is unknown, but the inference is robust to the unknown rate due to the use of the self-normalized statistic. In particular, this implies that our inference procedure does *not* require an estimation of the unknown index of stability α. Furthermore, it is not necessary to estimate the unknown slowly varying function either. These features are practical advantages of our proposed method. ▲

Remark 3 (Finite sample non-invertibility of other cluster-based resampling methods)**.** In comparison with the (conventional) subsampling, the score subsampling has two major advantages. First, as it does not require to recompute the inverse factor for each subsample, the score subsampling is computationally more efficient than the subsampling. Second, in finite samples, when regressors contain a cluster-specific binary treatment variable or other dummies variables that are highly correlated within a cluster, $\sum_{g \in \mathcal{B}_j} X_g' X_g$ can be often singular especially for small $b = |\mathcal{B}_j|$, and thus the subsampled OLS may not behave well for a non-negligible proportion of subsamples. This issue is also faced by other cluster-based resampling methods, such as the jackknife and bootstrap. In practice, several *ad hoc* 'fixes,' such as the use of generalized inverse or dropping such realizations, are employed. However, their theoretical implications remain unclear. Our cluster-robust score subsampling procedure avoids such an issue in a theoretically supported manner. ▲

4.2 Theoretical Properties

Section 4.2.1 establishes the asymptotic validity of the subsampling method under a fixed data generating process (DGP). Section 4.2.2 further extends it to the uniform validity over a broad class of DGPs.

4.2.1 Asymptotic Size Control under a Fixed DGP

The following theorem formally justifies the subsampling method under a fixed DGP.

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

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

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

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

and thus 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 branches into two cases. First, we focus on the pathological case with $\alpha < 2$. The statement for this case is presented as Lemma 1 in Appendix A.2, which is further proved in Appendix B.1. Appendix B.2 proves the statement for the case with $\alpha = 2$, and combines two cases to establish Theorem 2.

The limiting distribution, which is approximated by our proposed method of score subsampling, is not pivotal. It varies with two parameters: one is the index α of stability, and the other is p defined in (3.1), which measures the tail asymmetry of the distribution of $v'S_g$.

Choice of b: We close this section with discussions on the choice of b in practice. While the theory requires $b \to \infty$ and $b/G = o(1)$ as $G \to \infty$, a researcher needs to choose some value of b in a finite sample. We suggest to adapt the minimum volatility method (Algorithm 9.3.3 in Politis et al., 1999, Section 9.3.2) to our framework. Appendix A.4 provides a detailed algorithmic procedure that a practitioner can readily implement. We also employ this method to choose b in the numerical studies presented below.

4.2.2 Uniform Asymptotic Size Control

We now discuss the uniformity properties of the proposed score subsampling method. Without the uniformity, for any given G (regardless of size), there could exist a DGP, P_G, where the rejection probability under the null hypothesis fails to approach the desired level. Thus, the uniformity ensures reliable inference in finite samples, especially 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 $\mathbb{E}_P[S_g] = 0$, and there exist some $p \in [0, 1]$ and $\alpha \in [1 + \varepsilon, 2)$ such that

$$\lim_{t\to\infty} \frac{P(S_g > t)}{P(|S_g| > t)} = p, \quad \text{and} \tag{4.1}$$

$$P(|S_g| > t) = t^{-\alpha} L_P(t) \quad \text{as } t \to \infty \tag{4.2}$$

for an $L_P(\cdot)$ slowly varying at ∞ that can depend on $P = P_G$. 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. 790).

Proposition 2. *When $\alpha = 1$ in (4.2), the t-ratio $(\widehat{\delta} - \delta)/\widehat{\sigma}$ converges weakly to a nondegenerate limiting distribution only if S_g follows a (translation of) Cauchy distribution. 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: our proposed cluster score subsampling inference controls size uniformly over the set $\mathbf{P}(\varepsilon)$ if $\varepsilon > 0$. Note that the score subsampling coincides with (conventional) subsampling for sample means. Denote

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

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

Theorem 3 (Uniformity of the cluster score subsampling)**.** *For any $\varepsilon \in (0, 1]$, the confidence sets constructed based on cluster score subsampling 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*

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

A proof can be found in Appendix B.4. The proof utilizes the general results in Romano and Shaikh (2012) under high-level conditions together with our Lemma 2 in Appendix B.4. 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-normal ($\alpha_0 < 2$) 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-normal limiting distributions.

Finally, it is noteworthy that our uniform size control property exhibits resemblances to certain instances in the existing literature. An example is the AR(1) model presented in Example 1 of Andrews, Cheng, and Guggenberger (2020), where uniform size control persists across DGPs leading to either normal or non-normal 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 3 stands as the first theoretical result addressing the uniformity property of subsampling for statistical models that may exhibit potentially infinite variance.

4.3 Simulations

In this section, we present simulation studies to evaluate the finite sample performance of our proposed score subsampling method of CR inference in comparison with the 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 takes the value of one for $\lceil 0.2G \rceil$ clusters and zero for the remaining clusters $G - \lceil 0.2G \rceil$, where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. We draw cluster sizes $N_g \sim \lceil \text{Pareto}(1, \alpha) \rceil$ independently for $g \in \{1, \cdots, G\}$. For each $g \in \{1, \cdots, G\}$, we independently draw N_g-variate random vectors, $(\widetilde{X}_{g1j}, \cdots, \widetilde{X}_{gN_gj})' \sim \mathcal{N}(0, \Omega)$ for $j \in \{1, \cdots, K\}$ and $(\widetilde{U}_{g1}, \cdots, \widetilde{U}_{gN_g})' \sim \mathcal{N}(0, \Omega)$ in the baseline design, where Ω is an $N_g \times N_g$ variance-covariance matrix such that $\Omega_{ii} = 1$ for all $i \in \{1, \cdots, N_g\}$ and $\Omega_{ii'} = 1/2$ whenever $i \neq i'$. The controls are constructed by $X_{gij} = 0.2 F^{-1}_{\text{Beta}(2,2)} \circ \Phi(\widetilde{X}_{gij})$, where $F_{\text{Beta}(2,2)}$ and

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

We vary values of the exponent parameter $\alpha \in \{1.1, 1.2, \cdots, 1.9, 2.0\}$ across sets of simulations. The regression coefficients are fixed to $(\theta_0, \theta_1, \theta_2, \cdots, \theta_{K+1})' = (1, 1, 1, \cdots, 1)'$ throughout, whereas the dimension K of covariates vary as $K \in \{0, 5, 10\}$. We set the sample size (i.e., the number of clusters) to $G = 50$ across sets of simulations, which is close to the number of states in the U.S. Each set of simulations consists of 5,000 Monte Carlo iterations.

Figure 2 illustrates the Monte Carlo coverage frequencies. The horizontal axis measures the value of α, and the vertical axis measures the coverage frequency. In the legend, 'SUB' (respectively, 'WCB', 'JACK' and 'CR1') indicates the score subsampling (respectively, wild cluster bootstrap, jackknife standard error with normal critical value, and CR1 standard error with normal critical value). The nominal coverage probability of 95% is indicated by the horizontal gray line at 0.95.

When α is small, say $\alpha < 1.6$, the score subsampling performs the best, followed by the jackknife, the WCB, and the CR1. When α is larger, say $\alpha > 1.6$, the score subsampling still performs the best, followed by the WCB, the jackknife, and the CR1. Overall, the score subsampling robustly yields the coverage frequencies closest to the nominal probability of 95% across various values of α. All the conventional methods suffer from sever under-coverage especially for small values of α.

5 Size-Adjusted Reweighting as the Second Reliable Solution

The score subsampling method introduced in the previous section is theoretically robust and sound. However, its practical implementation may pose challenges with existing statistical software such as Stata. To address this limitation, we introduce an alternative solution based on size-adjusted reweighting, which can be easily implemented using Stata.

We first present the proposed method without theoretical details in Section 5.1. Its theoretical support follows in Section 5.2. Section 5.3 presents simulation studies.

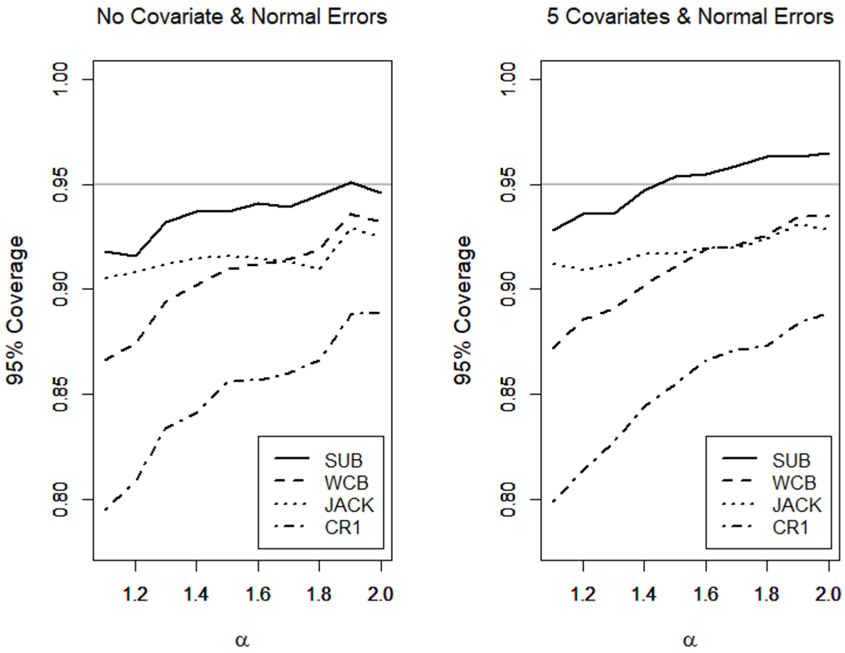

Figure 2: Monte Carlo coverage frequencies for the baseline design with normal errors. 'SUB' (respectively, 'WCB', 'JACK' and 'CR1') indicates the score subsampling (respectively, wild cluster bootstrap, jackknife standard error with normal critical value, and CR1 standard error with normal critical value). The nominal coverage probability of 95% is indicated by the horizontal gray line.

5.1 The Method

To accommodate arbitrarily non-ignorable clusters, we propose using the following size-adjusted least squares (SACR) method. Modify (2.1) and (2.2) by

$$\widehat{\theta}^{\mathrm{SACR}} = \left(\sum_{g=1}^{G} N_g^{-1} X_g' X_g\right)^{-1} \left(\sum_{g=1}^{G} N_g^{-1} X_g' Y_g\right) \qquad \text{and} \tag{5.1}$$

$$\widehat{V}_{\widehat{\theta}}^{\mathrm{SACR}} = a_G \left(\sum_{g=1}^{G} N_g^{-1} X_g' X_g\right)^{-1} \left(\sum_{g=1}^{G} N_g^{-2} \widehat{S}_g \widehat{S}_g'\right) \left(\sum_{g=1}^{G} N_g^{-1} X_g' X_g\right)^{-1}, \tag{5.2}$$

respectively, where $a_G \to 1$ almost surely. Note that a_G is stochastic under the current framework since N_g is random.

We provide a heuristic discussion about the SACR method and will discuss its theoretical properties in the next subsection. First, the SACR method is easy to implement using an existing Stata command and is compatible with nearly all regression-based Stata commands. Specifically, the SACR method can be readily implemented by

```
regress y x1 x2 ...  [aweight=1/N_g], cluster(cid)
```

in the Stata command line, and other commands may replace `regress y x1 x2`.

Second, to ensure that $\widehat{\theta}^{\mathrm{SACR}}$ is consistent for the estimand θ, we replace the conventional conditional mean independence assumption $\mathbb{E}[X_g U_g] = 0$ with the modified assumption $\mathbb{E}[N_g^{-1} X_g U_g] = 0$, accounting for the randomness of N_g in our framework. This assumption is no stronger than the conventional one, in which N_g is considered deterministic.

Third, compared to the conventional CR method, $\widehat{\theta}^{\mathrm{SACR}}$ assigns a weight of N_g^{-1} to the summands $X_g' X_g$ and $X_g' Y_g$ for each cluster g. This inverse cluster size mitigates the dominant effects of extremely large clusters. For example, with 51 states as clusters, $\widehat{\theta}^{\mathrm{SACR}}$ allocates equal weight to each state instead of each entity. The size adjustment transforms the intra-cluster summation $X_g' Y_g = \sum_{i=1}^{N_g} X_{gi} Y_{gi}$ into the intra-cluster average $N_g^{-1} X_g' Y_g = N_g^{-1} \sum_{i=1}^{N_g} X_{gi} Y_{gi}$, contributing to maintaining the asymptotic normality unlike the OLS.

5.2 Theoreical Properties

We now present the asymptotic properties of our SACR method. Again, let $\mathcal{P}$ be a generic class of models that generate the triplet (N_g, X_g, Y_g). For any $P \in \mathcal{P}$, we use the notations $\mathbb{E}_P$ for the expectation to emphasize its dependence on P. With these notations, we impose the following assumptions under slightly more structure than in Section 4.2.

Assumption 2 (Intra-Cluster)**.** There exists a $\delta > 0$ such that for all $P \in \mathcal{P}$, 1. $\mathbb{E}_P\left[\|X_{gi}\|^{4+4\delta} | N_g\right] \in [C_3, C_4]$ for $0 < C_1 < C_2 < \infty$ almost surely for all i and g. 2. $\mathbb{E}_P\left[\|X_{gi} U_{gi}\|^{2+2\delta} | N_g\right] \in [C_5, C_6]$ for $0 < C_3 < C_4 < \infty$ almost surely for all i and g.

Assumption 3 (Across-Cluster)**.** 1. (N_g, X_g, Y_g) is i.i.d. across g. 2. $\mathbb{E}_P[N_g^{-1} X_g' X_g]$ is non-singular.

Assumptions 2 and 3 are mild and sufficient for establishing the uniform asymptotic normality for our SACR method. Specifically, Assumption 2 requires that individual observations X_{gi}^2 and $X_{gi}U_{gi}$ have $2+\delta$ finite moments, which is common in the CR literature. Assumption 3 requires i.i.d.ness across clusters and a full rank condition, typically imposed in the literature when $G \to \infty$. Note that we do not assume any specific intra-cluster correlation structure or distribution of N_g, making it robust against (i) arbitrary intra-cluster correlations and (ii) non-ignorable clusters. These points are significant because point (i) is a primary motivation for using CR inference and point (ii) is a recent challenge for the conventional CR methods. Under Assumptions 2 and 3, the following theorem establishes the uniform asymptotic normality of our SACR method:

Theorem 4 (Uniformity of Size-Adjusted Reweighting)**.** *Suppose $\mathcal{P}$ is a set of DGPs that for each $P \in \mathcal{P}$, Assumptions 2 and 3 are satisfied. Then for any significance level $a \in (0,1)$ and any* $\dim(\theta)$*-vector $r \neq 0$, we have*

$$\sup_{P\in\mathcal{P}} \left| P\left(\frac{r'(\widehat{\theta}^{SACR}-\theta)}{\sqrt{r'\widehat{V}_{\widehat{\theta}}^{SACR} r/G}} > z_{1-a} \right) - a \right| = o(1).$$

as $G \to \infty$, where z_{1-a} denotes the $1-a$ quantile of the standard normal distribution.

A proof can be found in Appendix B.

This uniform validity is novel in the CR inference literature. We provide some discussions about this result. First, we note that Assumptions 2 and 3 can be regarded as sufficient conditions for Assumption 1, provided the score is redefined. Specifically, consider the cluster-level score $\widetilde{S}_g = N_g^{-1}\sum_{i=1}^{N_g} X_{gi}U_{gi}$. Assumptions 2 and 3 ensure that $\widetilde{S}_g$ has a finite second moment, thereby allowing the central limit theorem to apply as $G \to \infty$. Consequently, $\widetilde{S}_g$ belongs to the domain of attraction of stable laws with a stability index $\alpha = 2$. The structure of our SACR method enables us to derive more intuitive primitive conditions.

Second, since the asymptotic properties of the SACR method are derived under a different set of conditions, its uniformity is not directly comparable to that of the subsampling method studied in Section 4.2.2. However, under the environment of Theorem 4, the uniformity of SACR-based statistic appears more general, as it maintains uniform size control even when the distribution of cluster size N_g lacks a first moment—unlike the score subsampling method. As discussed in the introduction, the literature on urban economics and economic geography has established theoretical results indicating that city sizes follow Zipf's law with a unit exponent (e.g., Gabaix, 1999), which leads to the issue of the nonexistence of the first moment. Therefore, we recommend SACR for such applications.

5.3 Simulation Studies

In this section, we present simulation studies to evaluate the finite sample performance of the SACR methods in comparison with the conventional CR methods. The data-generating design is largely the same as that described in Section 4.3, but is redefined below for completeness. 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 takes the value of one for $\lceil 0.2G \rceil$ clusters and zero for the remaining clusters $G - \lceil 0.2G \rceil$, where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. We draw cluster sizes $N_g \sim \lceil 10 \cdot \text{Pareto}(1, \alpha) \rceil$ independently for $g \in \{1, \cdots, G\}$. For each $g \in \{1, \cdots, G\}$, we independently draw N_g-variate random vectors, $(\widetilde{X}_{g1j}, \cdots, \widetilde{X}_{gN_g j})' \sim \mathcal{N}(0, \Omega)$ for $j \in \{1, \cdots, K\}$ and $(\widetilde{U}_{g1}, \cdots, \widetilde{U}_{gN_g})' \sim \mathcal{N}(0, \Omega)$, where Ω is an $N_g \times N_g$ variance-covariance matrix such that $\Omega_{ii} = 1$ for all $i \in \{1, \cdots, N_g\}$ and $\Omega_{ii'} = 1/2$ whenever $i \neq i'$. The controls are constructed by $X_{gij} = 0.2 F^{-1}_{\text{Beta}(2,2)} \circ \Phi(\widetilde{X}_{gij})$, where $F_{\text{Beta}(2,2)}$ and Φ denote the CDFs of the Beta$(2, 2)$ and standard normal distributions, respectively. The errors are heteroskedastically constructed by $U_{gi} = 0.2 \widetilde{U}_{gi}$ if $T_g = 0$ and $U_{gi} = \widetilde{U}_{gi}$ if $T_g = 1$.

We vary values of the exponent parameter $\alpha \in \{1, 2, 4\}$ across sets of simulations. The regression coefficients are fixed to $(\theta_0, \theta_1, \theta_2, \cdots, \theta_{K+1})' = (1, 1, 1, \cdots, 1)'$ throughout, whereas the dimension K of covariates vary as $K \in \{0, 1, 5\}$. We set the sample size (i.e., the number of clusters) to $G = 50$ across sets of simulations, which is close to the number of states in the U.S. Each set of simulations consists of 10,000 Monte Carlo iterations.

In addition to the OLS (2.1) with the CR variance estimator (2.2), we also implement the jackknife variance estimator defined by

$$\widehat{V}_{\widehat{\theta}}^{\text{CR,JACK}} = \sum_{g=1}^{G} \left(\widehat{\theta}_{-g}^{\text{OLS}} - \widehat{\theta}^{\text{OLS}}\right) \left(\widehat{\theta}_{-g}^{\text{OLS}} - \widehat{\theta}^{\text{OLS}}\right)', \tag{5.3}$$

where $\widehat{\theta}_{-g}^{\text{OLS}}$ denotes the leave-one-cluster-out OLS estimator defined by

$$\widehat{\theta}_{-g}^{\text{OLS}} = \left(\sum_{h \neq g} X_h' X_h\right)^{-1} \left(\sum_{h \neq g} X_h' Y_h\right).$$

Likewise, the size-adjusted version of jackknife is

$$\widehat{V}_{\widehat{\theta}}^{\text{SACR,JACK}} = \sum_{g=1}^{G} \left(\widehat{\theta}_{-g}^{\text{SACR}} - \widehat{\theta}^{\text{SACR}}\right) \left(\widehat{\theta}_{-g}^{\text{SACR}} - \widehat{\theta}^{\text{SACR}}\right)',$$

where

$$\widehat{\theta}_{-g}^{\text{SACR}} = \left(\sum_{h \neq g} N_h^{-1} X_h' X_h\right)^{-1} \left(\sum_{h \neq g} N_h^{-1} X_h' Y_h\right)$$

denotes the leave-one-cluster-out SACR estimator.

Figures 3–4 draw Q-Q plots under $\alpha = 2$ and $\alpha = 1$, respectively, of the self-normalized statistics:

$$\begin{aligned}
&\text{(CR)} && \left(\widehat{\theta}_1^{\text{OLS}} - \theta_1\right) \Big/ \sqrt{\widehat{V}_{\widehat{\theta},11}^{\text{CR}}}, \\
&\text{(CR Jackknife)} && \left(\widehat{\theta}_1^{\text{OLS}} - \theta_1\right) \Big/ \sqrt{\widehat{V}_{\widehat{\theta},11}^{\text{CR,JACK}}}, \\
&\text{(SACR)} && \left(\widehat{\theta}_1^{\text{SACR}} - \theta_1\right) \Big/ \sqrt{\widehat{V}_{\widehat{\theta},11}^{\text{SACR}}}, \qquad \text{and} \\
&\text{(SACR Jackknife)} && \left(\widehat{\theta}_1^{\text{SACR}} - \theta_1\right) \Big/ \sqrt{\widehat{V}_{\widehat{\theta},11}^{\text{SACR,JACK}}}.
\end{aligned}$$

For these figures, we focus on the case with $K = 0$. In each figure, the dashed line indicates the 45° line, and the solid line indicates the fitted line.

Observe that the self-normalized statistics based on the conventional CR methods suffer from farther deviation away from the theoretical quantiles, whereas those based on the SACR methods more precisely follow the theoretical quantiles. This observation is true for both the analytic standard error estimator and the jackknife estimator. The deviations for the

$\alpha = 2$

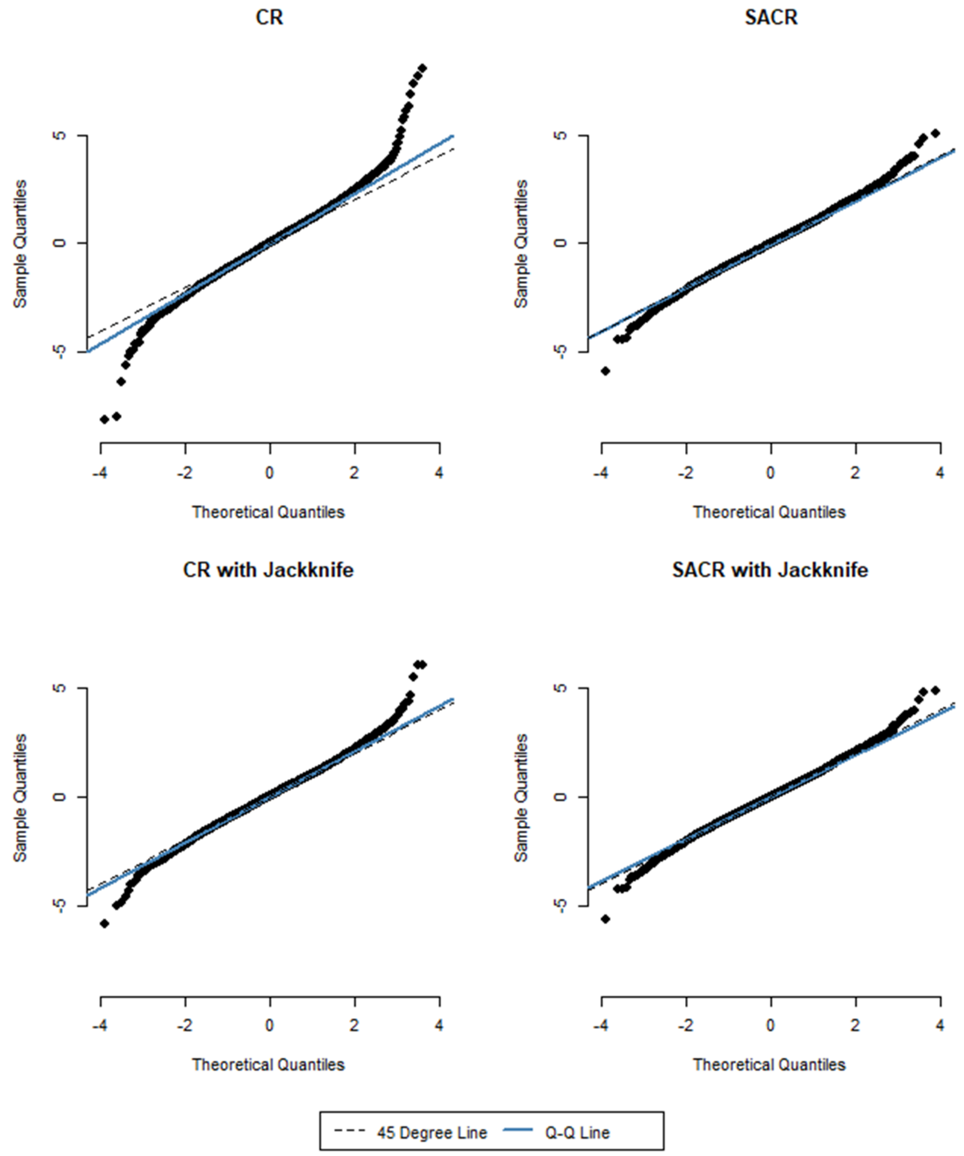

Figure 3: Q-Q plots for the self-normalized statistics under $\alpha = 2$. The dashed line indicates the 45° line, and the solid line indicates the fitted line. The left column shows the results for the conventional CR methods, while the right column shows the results for the SACR methods. The top row shows the results with analytic standard errors, while the bottom row shows the results with jackknife estimators.

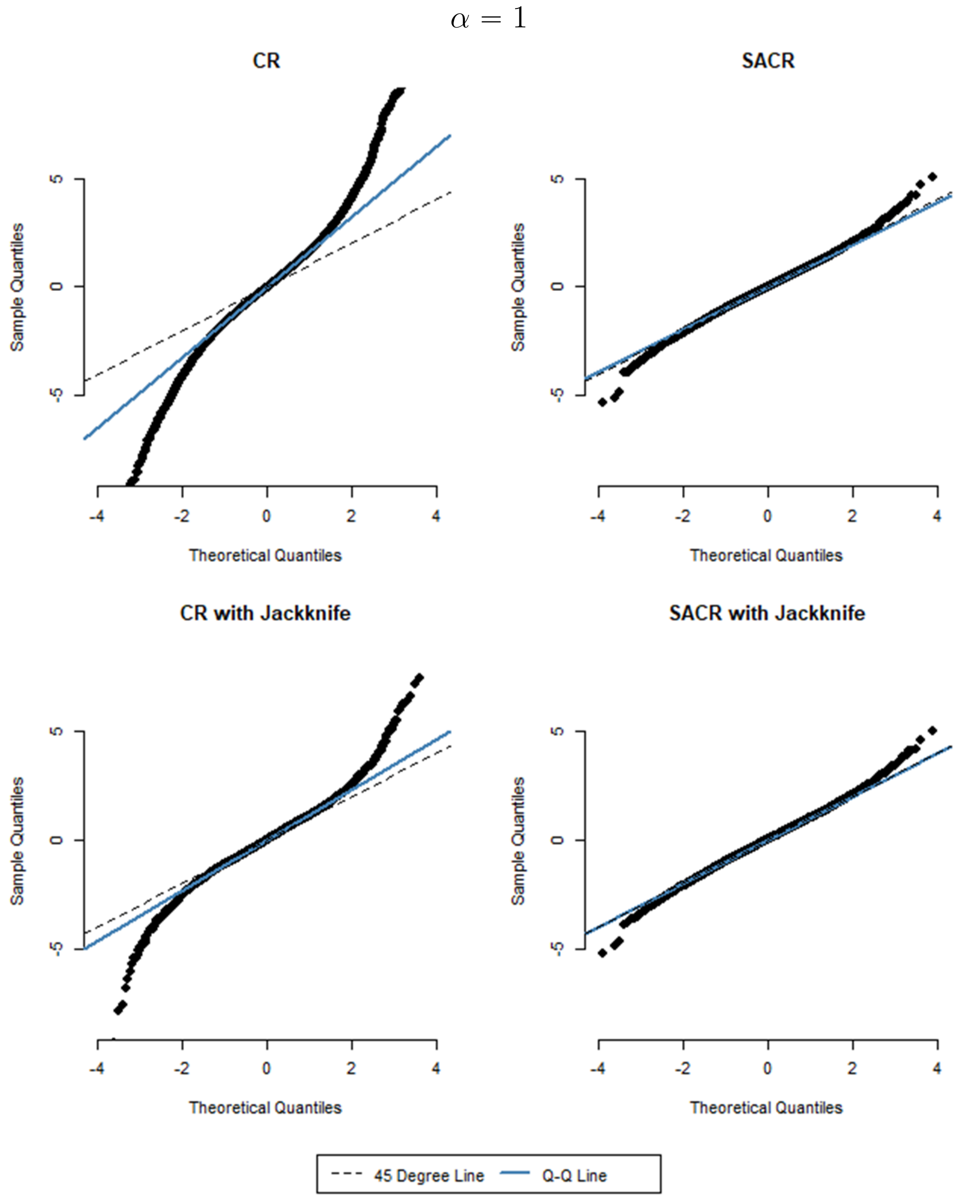

Figure 4: Q-Q plots for the self-normalized statistics under $\alpha = 1$. The dashed line indicates the 45° line, and the solid line indicates the fitted line. The left column shows the results for the conventional CR methods, while the right column shows the results for the SACR methods. The top row shows the results with analytic standard errors, while the bottom row shows the results with jackknife estimators.

		Conventional CR			New SACR		
		MSE	Rejection (level = 0.050)		MSE	Rejection (level = 0.050)	
K	α	$\widehat{\theta}_1^{\text{OLS}}$	$\widehat{V}^{\text{CR}}_{\widehat{\theta},11}$	$\widehat{V}^{\text{CR,JACK}}_{\widehat{\theta},11}$	$\widehat{\theta}_1^{\text{SACR}}$	$\widehat{V}^{\text{SACR}}_{\widehat{\theta},11}$	$\widehat{V}^{\text{SACR,JACK}}_{\widehat{\theta},11}$
0	4	0.057	(0.095)	[0.072]	0.054	(0.088)	[0.067]
	2	0.077	(0.141)	[0.088]	0.055	(0.086)	[0.069]
	1	0.144	(0.272)	[0.106]	0.053	(0.073)	[0.068]
1	4	0.058	(0.096)	[0.073]	0.054	(0.088)	[0.068]
	2	0.074	(0.136)	[0.085]	0.054	(0.087)	[0.070]
	1	0.138	(0.273)	[0.108]	0.053	(0.074)	[0.070]
5	4	0.057	(0.094)	[0.065]	0.054	(0.082)	[0.063]
	2	0.071	(0.130)	[0.082]	0.053	(0.079)	[0.064]
	1	0.121	(0.254)	[0.101]	0.053	(0.070)	[0.068]

Table 4: Simulation results based on 10,000 Monte Carlo iterations. Displayed are the mean square error (MSE), the rejection frequencies based on the analytic standard error estimation in round brackets with the nominal probability of 0.05, and the rejection frequencies based on the jackknife standard error estimation in square brackets with the nominal probability of 0.05. The first three columns show the results for the conventional CR methods, and the last three columns show the results for the SACR method.

conventional CR methods further exacerbate as the tail of N_g becomes heavier, as in the transition from $\alpha = 2$ (in Figure 3) to $\alpha = 1$ (in Figure 4). These results are consistent with the non-normality of the conventional CR methods as discussed in Section 3, as well as the guaranteed normality of the SACR methods as shown in Theorem 4.

Table 4 summarizes simulation results. Displayed are the mean square error (MSE), the rejection frequencies based on the analytic standard error estimation in round brackets, and the rejection frequencies based on the jackknife standard error estimation in square brackets. The nominal probability of rejection is set to 0.05 throughout. The first three columns show the results for the conventional CR methods. The last three columns show the results for the SACR methods.

We find the following three observations in these simulation results. First, focus on the MSE. While the MSE of the OLS estimator $\widehat{\theta}_1^{\mathrm{OLS}}$ exponentially blows up as α decreases, the MSE of the SACR estimator $\widehat{\theta}_1^{\mathrm{SACR}}$ remains stable as α varies. Second, consider the rejection frequencies reported in the round brackets based on the analytic standard error estimators, $\widehat{V}_{\widehat{\theta},11}^{\mathrm{CR}}$ and $\widehat{V}_{\widehat{\theta},11}^{\mathrm{SACR}}$. While the rejection frequencies for the conventional CR method based on $\widehat{V}_{\widehat{\theta},11}^{\mathrm{CR}}$ blows up as α decreases, the rejection frequencies for the SACR method based on $\widehat{V}_{\widehat{\theta},11}^{\mathrm{SACR}}$ remain stable and closer to the nominal rejection probability of 0.05 as α varies. Third, consider the rejection frequencies reported in the square brackets based on the jackknife standard error estimators, $\widehat{V}_{\widehat{\theta},11}^{\mathrm{CR,JACK}}$ and $\widehat{V}_{\widehat{\theta},11}^{\mathrm{SACR,JACK}}$. While these jackknife standard error estimators deliver more desirable rejection frequencies than the analytic standard error estimators for each of the conventional CR methods and the new SACR method, we continue to observe the same qualitative pattern as in the case of the analytic estimators. Namely, while the rejection frequencies for the conventional CR method based on $\widehat{V}_{\widehat{\theta},11}^{\mathrm{CR,JACK}}$ blows up as α decreases, the rejection frequencies for the SACR method based on $\widehat{V}_{\widehat{\theta},11}^{\mathrm{SACR,JACK}}$ remain stable and closer to the nominal rejection probability of 0.05 as α varies.

From these observations, it seems more desirable to use the SACR method over the conventional CR methods for estimation accuracy as well as robust inference.

6 Summary

Conventional methods for cluster-robust inference often fail to provide consistent results when faced with unignorably large clusters. In this paper, we formalize this limitation by deriving a necessary and sufficient condition for consistency. We find that 77% of empirical research articles published in the *American Economic Review* and *Econometrica* in 2020-2021 fail to satisfy this condition. To address this challenge, we propose two alternative solutions: (i) score subsampling and (ii) size-adjusted cluster-robust (SACR) estimation. Both methods ensure uniform size control across a wide range of data-generating processes where conventional methods fall short.

The first approach retains the original least squares estimator and approximates its non-standard asymptotic distribution. The second one changes the estimator to a weighted least squares, which is is easily implementable in Stata. Our simulation studies confirm the reliability and effectiveness of these methods, highlighting their practical utility in addressing the limitations of existing cluster-robust inference techniques.

Appendix

A Omitted Details

This appendix section collectes 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 5 (de la Peña et al., 2009, Theorem 2.24)**.** *Suppose $\alpha < 2$. Then, ξ_g belongs to the domain of attraction of an α-stable distribution if and only if*

$$P(|\xi_g| > t) = t^{-\alpha} L(t) \qquad \textit{and}$$
$$\lim_{t \to \infty} \frac{P(\xi_g > t)}{P(|\xi_g| > t)} = p, \quad p \in [0,1],$$

for some slowly varying function $L(\cdot)$.

The first condition 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 in this alternative characterization imposes a mild restriction on the existence of limiting ratios of one-sided tail probabilities over the two-sided tail probability; it rules out some irregular, infinitely oscillating type situations such that these limiting ratios do not exist. This condition only imposes restrictions in the limit and accommodates a wide range of tail behaviors as p are permitted to be either 0 or 1.

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

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

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

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

and thus

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

A proof is provided in Appendix B.1.

Remark 4 (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 5 (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 (B.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

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

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

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

required in the literature (e.g. Assumption 2 in Hansen and Lee 2019) for normal approximation.[8]

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 6 (On CR standard error estimation)**.** The test statistic we consider is the standard t-statistic used in the literature. Its denominator consists of a CR standard error without imposing a null hypothesis. When $\alpha < 2$, the asymptotic variance does not exist, and nor is this "standard error" consistent but remains random asymptotically. This is similar in spirit to the fixed-b asymptotics (e.g., Kiefer and Vogelsang, 2002) in the literature of long-run variance estimation, although the underlying theory is completely different as the fixed-b asymptotics crucially relies on normal approximation and the functional central limit theo-

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

rem. 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) for example. ▲

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 5. A remaining and potentially more interesting question is whether one can prove its validity using an alternative argument. The following result suggests that such efforts are ill-fated when $\alpha < 2$.

For simplicity of illustration, consider the case of a univariate regression with only the intercept, i.e. a cluster sampled mean $\widehat{\theta} = N^{-1}\sum_{g=1}^{G}\sum_{i=1}^{N_g} Y_{gi}$ with the cluster specific population mean normalized to $\theta = \mathbb{E}\left[\sum_{i=1}^{N_g} Y_{gi}\right] = 0$ without loss of generality. Suppose that the parameter 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 6 (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 B.6

A.4 Choosing the Number b of Cluster Subsamples

For the choice of b in practice, we adapt the minimum volatility method (Algorithm 9.3.3 in Politis et al., 1999, Section 9.3.2) to our framework of cluster-robust inference.

For subsampling to be valid, b needs to grow with the number G of clusters but at a slower rate. If b is too close to G, then all the subsampled t-statistics will be almost identical to the full-sample t-statistic, resulting in a subsampling distribution being too tight and thus in under-coverage by confidence intervals. On the other hand, if b is too small, then the subsampled t-statistics will be noisy and can result in either under-coverage or over-coverage. Thus, intuitively, we wish to select a b that is in a stable range for the test statistic. The following algorithm formalizes such an idea.

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

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

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

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

Remark 7. As pointed out by Romano and Wolf (1999, Section 11.5), empirical bootstrap is not valid in the presence of heavy-tailed distributions. Thus, the common calibration method for the choice of subsampling block size cannot be used in our setup. ▲

A.5 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

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

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.1). The independence within cluster implies that conditional on $\{N_g\}_{g=1}^G$,

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

for the numerator of (A.1). 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

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

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

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, $S_g = \sum_{i=1}^{N_g} Y_{gi} = N_g Y_g$, yielding that

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

The denominator still converges to $\sqrt{\mathbb{E}\left[N_g\right]}$. For the numerator, since $N_g Y_g$ 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 $O_p\left(1\right)$. 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-normal. 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

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

Following the same arguments as those in Case (ii), the first item above is asymptotically non-normal (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.2), Case (i) derives that $G^{-1/2}\sum_{g=1}^G\sum_{i=1}^{N_g} U_{gi} = O_p(1)$. The non-normal 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 $\sup_g N_g/N \to 0$ 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 Mathematical Proofs

This section collects all the mathematical proofs. The order in which the proofs appear below differs from the order in which the corresponding statements appear. Namely, the proof of Theorem 2 uses Lemma 1, and hence we present the proof of Lemma 1 before the proof of Theorem 2. Furthermore, the proof of Theorem 1 uses Lemma 1 and Theorem 2, and hence we present the proofs of Lemma 1 and Theorem 2 before the proof of Theorem 1. Proofs for all the remaining theorems are presented in the order of the appearance of their correponding results, that is, Theorem 3 (the uniform validity of the subsampling method), Theorem 4 (the uniform validity of the SACR method), and Theorem 6 (the failure of wild bootstrap). Some technical lemmas are relegated to Appendix C.

B.1 Proof of Lemma 1

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

Step 1. Recall that

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

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

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

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

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

as $G \to \infty$, where

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

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

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

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

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

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

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 (B.3) implies the consistency

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

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

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

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

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

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

holds, where the second inequality follows from the Cauchy-Schwarz inequality and the stochastic orders are due to Equations (B.3), (B.4), and (B.5). Second, similar lines of calculations yield

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

We have now established that

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

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

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

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

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

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

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

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

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

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

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

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

B.2 Proof of Theorem 2

Proof of Theorem 2. 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 (B.6), it holds that $SN_2(\theta) = SN_1(\theta) + o_p(1)$. Finally, for the validity of the subsampling procedure, we now invoke Theorem 2.2.1 in Politis et al. (1999) and note that the limiting distribution is normal and hence continuous. □

B.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 (B.6) satisfy $SN_2(\theta) = SN_1(\theta) + o_p(1)$, as shown in the proofs for Lemma 1 and Theorem 2. □

B.4 Proof of Theorem 3

Proof of Theorem 3. 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 3, we shall derive the asymptotic distribution for the following self-normalized sums of S_g:

$$R_{1G} := \frac{\sum_{g=1}^{G} S_g}{\sqrt{\sum_{g=1}^{G} S_g^2}} \quad \text{and} \quad R_{2G} := \frac{\widehat{\delta} - \delta}{\widehat{\sigma}} = \frac{\sum_{g=1}^{G} S_g}{\sqrt{\sum_{g=1}^{G} \widehat{S}_g^2}}. \tag{B.6}$$

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 C 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

$$\sup_{P \in \mathbf{P}} \sup_{x \in \mathbb{R}} |J_b(x, P) - J_G(x, P)| \to 0 \tag{B.7}$$

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 (B.7) 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

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

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, 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 $\{Z_k\}_k$. 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 3. □

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

$$\left(\frac{1}{A_{Mn}} \sum_{g=1}^{n} S_g(\alpha_M, p_M), \frac{1}{A_{Mn}^2} \sum_{g=1}^{n} S_g^2(\alpha_M, p_M) \right)$$
$$\xrightarrow{d} \left(\sum_{k=1}^{\infty} \{\epsilon_k(p_M) Z_k(\alpha_M) - (2p_M - 1)\mathbb{E}[Z_k(\alpha_M)\mathbb{1}(Z_k(\alpha_M) < 1)]\}, \sum_{k=1}^{\infty} Z_k^2(\alpha_M) \right) = (S_M, V_M)$$

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 $X_{Mn} \overset{a.s.}{\to} X_M$ for each M and as $n \to \infty$, and $X_M \overset{a.s.}{\to} X$ 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 C, 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 C. □

B.5 Proof of Theorem 4

Proof of Theorem 4. Without loss of generality, assume that X_{gi} is a scalar. We show the claim that for any sequence of DGPs $(P_G)_{G=1}^{\infty} \subset \mathcal{P}$, it holds that

$$\frac{(\widehat{\theta}^{\text{SACR}} - \theta)}{\sqrt{\widehat{V}_{\widehat{\theta}}^{\text{SACR}}/G}} \xrightarrow{d} \mathcal{N}(0,1).$$

Given such a claim, the result follows from a standard assume toward a contradiction argument that can be found in, e.g. Theorem 11.4.5 in Lehmann and Romano (2005). Explicitly, if the statement of the theorem fails, one could extract a subsequence, with an abuse of notation still denoted as $(P_G)_{G=1}^{\infty}$, such that for an $a' \in (0,1)$

$$P_G\left(\frac{(\widehat{\theta}^{\text{SACR}}-\theta)}{\sqrt{\widehat{V}_{\widehat{\theta}}^{\text{SACR}}/G}} > z_{1-a}\right) \to a' \neq a,$$

which violates the claim, a contradiction.

To prove the claim, fix any sequence of DGPs $(P_G)_{G=1}^{\infty} \subset \mathcal{P}$. For $(P_G)_{G=1}^{\infty}$, we first establish the asymptotic normality of $\widehat{\theta}^{\text{SACR}}$ along this sequence of DGPs

$$\frac{(\widehat{\theta}^{\text{SACR}}-\theta)}{\sqrt{V_{\widehat{\theta}}^{\text{SACR}}}} \xrightarrow{d} \mathcal{N}(0,1).$$

by invoking a CLT for triangular arrays (e.g., Lehmann and Romano, 2005, Lemma 11.4.1) and a WLLN for triangular arrays (e.g., Lehmann and Romano, 2005, Lemma 11.4.2).

Given the row-wise i.i.d. triangular array assumption, it suffices to establish that for a $\delta > 0$

$$\lim_{\lambda\to\infty}\limsup_{G\to\infty}\mathbb{E}_P[|N_g^{-1}X_g'X_g|\mathbb{1}\{|N_g^{-1}X_g'X_g| > \lambda\}] = 0 \tag{B.8}$$

$$\lim_{\lambda\to\infty}\limsup_{G\to\infty}\mathbb{E}_P[|N_g^{-1}X_g'U_g|^2\mathbb{1}\{|N_g^{-1}X_g'X_g| > \lambda\}] = 0. \tag{B.9}$$

Without loss of generality and for ease of writing, we show them for the case with a scalar X_{gi}. For (B.8), note that for each fixed $\lambda > 0$ Assumption 2.1 yields that for some finite constant C

$$\begin{aligned}\mathbb{E}_P[|N_g^{-1}X_g'X_g|\mathbb{1}\{|N_g^{-1}X_g'X_g| > \lambda\}] \leqslant& \frac{1}{\lambda^{\delta}}\mathbb{E}_P[|N_g^{-1}X_g'X_g|^{1+\delta}] \\ =& \frac{1}{\lambda^{\delta}}\mathbb{E}_P\left[\mathbb{E}_P\left[\left|\frac{1}{N_g}\sum_{i=1}^{N_g}X_{gi}^2\right|^{1+\delta}|N_g\right]\right] \\ \leqslant& \frac{1}{\lambda^{\delta}}\mathbb{E}_P\left[\frac{1}{N_g}\sum_{i=1}^{N_g}\mathbb{E}_P\left[\left|X_{gi}^2\right|^{1+\delta}|N_g\right]\right] < \frac{1}{\lambda^{\delta}}C\end{aligned}$$

where the second inequality follows from Jensen's inequality and the last follows from $\mathbb{E}_P\left[\left|X_{gi}^2\right|^{1+\delta}|N_g\right] < C$.

For (B.9), for each fixed $\lambda > 0$, Assumption 2.2 yields that for some finite constant C', it holds that

$$\begin{aligned}
&\mathbb{E}_P[|N_g^{-1}X_g'U_g|^2\mathbb{1}\{|N_g^{-1}X_g'U_g| > \lambda\}] \\
\leqslant&\frac{1}{\lambda^\delta}\mathbb{E}_P[|N_g^{-1}X_g'U_g|^{2+\delta}] \\
=&\frac{1}{\lambda^\delta}\mathbb{E}_P\left[\mathbb{E}\left[\left|\frac{1}{N_g^2}\sum_{i=1}^{N_g}\sum_{i'=1}^{N_g}X_{gi}'U_{gi}U_{gi'}X_{gi'}\right|^{1+\delta/2}|N_g\right]\right] \\
\leqslant&\frac{1}{\lambda^\delta}\mathbb{E}_P\left[\frac{1}{N_g^2}\sum_{i=1}^{N_g}\sum_{i'=1}^{N_g}\mathbb{E}\left[\left|X_{gi}'U_{gi}U_{gi'}X_{gi'}\right|^{1+\delta/2}|N_g\right]\right] < \frac{1}{\lambda^\delta}C'
\end{aligned}$$

following Jensen's inequality and as

$$\begin{aligned}
\mathbb{E}_P\left[\left|X_{gi}'U_{gi}U_{gi'}X_{gi'}\right|^{1+\delta/2}|N_g\right] \leqslant&\sqrt{\mathbb{E}_P\left[\left|X_{gi}'U_{gi}\right|^{2+\delta}|N_g\right]\mathbb{E}_P\left[\left|X_{gi'}'U_{gi'}\right|^{2+\delta}|N_g\right]} \\
\leqslant&\mathbb{E}_P\left[\left|X_{gi}'U_{gi}\right|^{2+\delta}|N_g\right] < C'.
\end{aligned}$$

Now we show the ratio consistency of the variance estimator along $(P_G)_{G=1}^\infty$

$$G\widehat{V}_{\hat{\theta}}^{\text{SACR}}/V^{\text{SACR}} \xrightarrow{p} 1.$$

Note that

$$\begin{aligned}
&\mathbb{E}_P[|N_g^{-1}X_g'X_g|^2\mathbb{1}\{|N_g^{-1}X_g'X_g| > \lambda\}] \\
\leqslant&\frac{1}{\lambda^{2\delta}}\mathbb{E}_P[|N_g^{-1}X_g'X_g|^{2+2\delta}] \\
\leqslant&\frac{1}{\lambda^{2\delta}}\mathbb{E}_P\left[\mathbb{E}_P\left[\left|\frac{1}{N_g}\sum_{i=1}^{N_g}X_{gi}^2\right|^{2+2\delta}|N_g\right]\right] \\
\leqslant&\frac{1}{\lambda^{2\delta}}\mathbb{E}_P\left[\frac{1}{N_g}\sum_{i=1}^{N_g}\mathbb{E}_P\left[\left|X_{gi}\right|^{4+4\delta}|N_g\right]\right] < \frac{1}{\lambda^\delta}C
\end{aligned}$$

which converges to zero as $\lambda \to \infty$. Hence, the convergence $G^{-1}\sum_{g=1}^G N_g^{-1}X_g'X_g \xrightarrow{p} \mathbb{E}_P[N_g^{-1}X_g'X_g]$ follows from WLLN for triangular arrays given (B.8) and Assumption 3.1. Following the assumption on a_G, it holds that $a_G \to 1$ almost surely. (Recall that a_G is stochastic in general under our framework in which N_g is random.) Then, it remains to show that along $(P_G)_G$, it holds that

$$G^{-1}\sum_{g=1}^{G} N_g^{-2}\widehat{S}_g\widehat{S}_g' \xrightarrow{p} \mathbb{E}_P[N_g^{-2}X_g'U_gU_g'X_g].$$

To this end, we first write

$$\begin{aligned}\widehat{S}_g &= \sum_{i=1}^{N_g} X_{gi}\widehat{U}_{gi}\\ &= \sum_{i=1}^{N_g} X_{gi}U_{gi} - \sum_{i=1}^{N_g} X_{gi}^2\,(\widehat{\theta}-\theta)\\ &= S_g - \Xi_g\,(\widehat{\theta}-\theta),\end{aligned}$$

where $\Xi_g = \sum_{i=1}^{N_g} X_{gi}^2$. Then, we have

$$\begin{aligned}&G^{-1}\sum_{g=1}^{G} N_g^{-2}\widehat{S}_g\widehat{S}_g'\\ =&\,G^{-1}\sum_{g=1}^{G} N_g^{-2}S_g^2 - 2G^{-1}\sum_{g=1}^{G} N_g^{-2}S_g\Xi_g(\widehat{\theta}-\theta) + G^{-1}\sum_{g=1}^{G} N_g^{-2}\Xi_g^2(\widehat{\theta}-\theta)^2\\ =&\,\mathbb{E}_P[N_g^{-2}S_g^2] - 2\mathbb{E}_P[N_g^{-2}S_g\Xi_g](\widehat{\theta}-\theta) + \mathbb{E}_P[N_g^{-2}\Xi_g^2](\widehat{\theta}-\theta)^2 + o_p(1)\\ =&\,\mathbb{E}_P[N_g^{-2}S_g^2] \;+\; o_p(1),\end{aligned}$$

by WLLN for triangular arrays given (B.8), (B.9), Assumption 2, Assumption 3, and the consistency of $\widehat{\theta}$ from the first part of this proof. □

B.6 Proof of Theorem 6

Proof of Theorem 6. 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

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

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

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

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

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

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 Auxiliary Lemmas

Let $X^{(G)} = (X_1, ..., X_G)$ be a sequence of i.i.d. random variables with distribution $P \in \mathbf{P}$ and let the distribution of a real-valued root $R_G = R_G(X^{(G)}, P)$ under P be denoted by $J_G(x, P)$. In addition, for a subsample size $b = b_G < G$ such that $b = o(G)$, define $B_G = \binom{G}{b}$. For $j = 1, ..., B_G$, let $X^{G,(b),j}$ denote the j-th subsample of size b. Define

$$\begin{aligned} L_G(x, P) =& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1}\{R_b(X^{G,(b),j}, P) \leqslant x\} \quad \text{and} \\ \widehat{L}_G(x) =& \frac{1}{B_G} \sum_{j=1}^{B_G} \mathbb{1}\{\widehat{R}_b(X^{G,(b),j}) \leqslant x\}, \end{aligned}$$

where $\widehat{R}_b$ is a feasible estimator of R_b, which depends on the unknown P.

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)**.** *Under the current setup,*

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

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 taken from 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.

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.

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.

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.

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.

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.

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

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

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

Embrechts, P. 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.

Gabaix, X. (1999): "Zipf's law for cities: an explanation," *The Quarterly journal of economics*, 114, 739–767.

——— (2009): "Power laws in economics and finance," *Annu. Rev. Econ.*, 1, 255–294.

——— (2016): "Power laws in economics: An introduction," *Journal of Economic Perspectives*, 30, 185–206.

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.*

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

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

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

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.

Lehmann, E. and J. P. Romano (2005): "Testing Statistical Hypotheses," *Springer Texts in Statistics.*

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.

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

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

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

Reichardt, C. S. and H. F. Gollob (1999): "Justifying the use and increasing the power of at test for a randomized experiment with a convenience sample." *Psychological Methods*, 4, 117.

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.

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

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

Sasaki, Y. and Y. Wang (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.