Comparing variance estimators: a test-based relative-efficiency approach *

Samuel P. Engle†

Job Market Paper

University of Wisconsin-Madison

Department of Economics

This version: October 16, 2021
Please find the current version here

Abstract

When constructing Wald tests, consistency is the key property required for the variance estimator. This property ensures asymptotic validity of Wald tests and confidence intervals. Classical efficiency comparisons indicate all consistent variance estimators lead to equivalent tests. This paper derives asymptotic normality of the Wald test statistic under fixed alternatives. These asymptotic distributions generally differ when different variance estimators are used. A simple relative efficiency measure leads to several new conclusions: when analyzing heavy-tailed data there are tradeoffs between point estimation and variance estimation, smooth kernels should be used when estimating the asymptotic variance in quantile regression, and conservative cluster-robust inference leads to efficiency loss. Simulation evidence indicates these results provide good finite-sample approximations. An application demonstrates how researchers considering different cluster levels can evaluate the costs associated with conservative clustering.

*I am grateful for the encouragement, advice, and, especially, patience I have received from Jack Porter, Bruce Hansen, Mikkel Sølvsten, and Harold Chiang. I also thank Xiaoxia Shi, Ken West, Bo Honore, Jim Powell, Eric Auerbach, Yuya Sasaki, Kei Hirano, Anson Zhou, Anna Trubnikova, Annie Lee, John Stromme, and past seminar participants at the University of Wisconsin-Madison for their helpful insights and comments which have shaped the paper in its present form. I gratefully acknowledge the financial support I have received from the Alice S. Gengler Dissertation Fellowship. All remaining errors are my own.

†email: sengle2@wisc.edu website: samuelpengle.com

1 Introduction

Much of empirical work in economics follows a three step recipe: estimate the parameter of interest, estimate the asymptotic variance, then construct a test statistic or confidence interval to answer the research question of interest. The first step is generally treated differently than the other two; while discussions on parameter estimation often focus on efficiency, the dialogue around variance estimation and testing typically focuses on robustness to misspecification. In this paper we demonstrate that this focus on robustness ignores meaningful implications for efficiency in the variance estimation step. The resulting asymptotic theory provides a theoretical foundation for several common "folk" theorems in applied work.

This paper makes three key contributions to the econometric literature on hypothesis testing. First, we develop a large-sample theory of the behavior of test statistics under fixed alternatives. Second, we connect the asymptotic distribution theory to a practical asymptotic relative efficiency measure. Third, we demonstrate that the convergence rate of a test statistic can differ from the convergence rate of the point estimate being used, and this has consequences in hypothesis testing.

The theory developed in this paper takes a different approach compared with the traditional local-asymptotic theory of Engle (1984), Newey and McFadden (1994), and van der Vaart (1998). That work finds that a broad class of tests statistics are found to have the same limiting distribution under local-alternatives. Our analysis is non-local, which leads to these equivalencies no longer holding in general. This allows for finer distinctions between testing procedures. In the case of Wald tests, local equivalence holds whenever the same estimator is used for two different tests, even if different consistent variance estimators are used. This equivalence breaks down in our asymptotic theory different estimators of the asymptotic variance are used. To compare these test statistics, we propose using an approximate asymptotic relative efficiency (ARE) measure which can be computed from the asymptotic distribution of the test statistic under a fixed alternative. Our approach contrasts with the local ARE of Pitman (1949), the most commonly used approach in econometrics. Under a fixed alternative, the Wald test statistic diverges to infinity, leading to power converging to 1. We evaluate a testing procedure by approximating the rate at which the test statistic diverges. The ARE measure we propose is an approximate version of the measure proposed in Hodges and Lehmann (1956) and is an extension of the approach in Hettmansperger (1972). There has been other recent work in econometrics on non-local ARE measures. Kim and Perron (2009) propose using an approximate version of Bahadur ARE (Bahadur (1967)) to test for structural breaks in time series. Canay and Otsu (2012) used Hodges-Lehmann ARE to assess the efficiency of generalized method of moments (GMM) and generalized empirical

likelihood tests of moments conditions. A benefit of our approach is broad applicability to testing problems most frequently encountered in empirical work.

Our distributional theory extends results in Bentkus et al. (2007), Omey and van Gulck (2009), and Shao and Zhang (2009). We establish the basic theory in the context of smooth GMM problems under i.i.d. sampling. We then show how to extend the results in several ways. These extensions serve as the basis for the comparison of variance estimators in this paper. In the case of quantile regression and heavy-tailed data, the asymptotic distribution of the test statistic is completely determined by the variance estimator under a fixed alternative. This implies that the asymptotic power properties of test statistics are determined by the variance estimator.

Within the i.i.d. setting, our first extension is to non-differentiable moment conditions. We focus on the linear conditional quantile regression model of Koenker and Bassett (1978). In this case, classic approaches to variance estimation involve estimators of the conditional density of the error term. We focus on the kernel density estimator of Powell (1991). In Kato (2012), the asymptotic distribution is derived for the kernel density estimator for the particular choice of a uniform kernel. The default choices in the `quantreg` package in `R` and when using the `qreg` function in Stata are the Gaussian and Epanechnikov kernel, respectively. We provide a first-order theoretical justification for this, by showing these estimators are more efficient relative to the uniform kernel.

The second extension we provide is for heavy-tailed data in linear instrumental variables (IV) models. Previous work has studied this environment with a focus on heavy tailed structural errors: the two-stage least absolute deviations estimator of Amemiya (1982) and Powell (1983), the rank estimator of Honoré and Hu (2004), the robust estimator in Sølvsten (2020), and a sequential testing procedure in Jiao (2019). Sasaki and Wang (2020) propose a test to ascertain whether GMM moments conditions have sufficient moments for asymptotic normality. We complement this work by demonstrating that there are also efficiency gains to be made in using more robust variance estimators in the presence of heavy-tailed data. In Müller (2020) it is noted that the convergence rate of test statistics to the null limit distribution is slower when the data are heavy-tailed. We show that under fixed alternatives there are also first-order implications for power. These points are empirically relevant. In a recent paper, Young (2021) argues that this issue is empirically relevant in IV models in applied work. Shephard (2020) considered testing in regression models with heavy-tailed regressors, focusing on accuracy of the normal approximation under the null. We show that the proposed procedure, adapted to the IV problem, could also lead to efficiency gains when the instruments have heavy tails. We also consider heavy-tailed first-stage errors. This is the first paper we are aware of that suggests first-stage outliers could negatively impact efficiency

of tests. We show that the Anderson-Rubin test (Anderson and Rubin (1949), Andrews et al. (2019)) is less robust than the Wald test when the first-stage errors have heavy tails.

The third extension we provide is to cluster-robust inference. The general framework we adopt is from Hansen and Lee (2019). Popularized in Bertrand et al. (2004), some recent work in econometrics has focused on the choice of cluster level. In Cameron and Miller (2015) it is argued that the coarsest cluster level should always be used. Abadie et al. (2017) presents a design-based approach to choosing the appropriate cluster level, along with some finite-sample results. MacKinnon et al. (2020b) provide a sequential testing procedure to detect the correct clustering level. We show that there is an unambiguous loss of efficiency when conservative clusters are used. Our results imply a method for researchers to conduct power analysis to see if the efficiency loss in their case is severe, or if there is little to be lost from the added robustness.

The rest of the paper proceeds as follows: we start by introducing the principles of our analysis in the context of two simple testing problems: hypothesis testing for means and medians. In Section 3, we provide a treatment of the distribution of Wald statistics in GMM settings, under fixed alternatives. In that section we provide extensions to heavy-tailed data settings, and the GMM setting with non-smooth moment conditions. These examples exhibit situations where the limit distribution is possibly non-normal, and the test statistic does not converge at a parametric rate under fixed alternatives. We then provide an important result for clustered data in Section 4. Simulations are provided in Section 5 to show the efficacy of the methods here in making finite-sample predictions. In Section 6, we apply our procedure to perform power analysis in the case of clustered sampling settings. A summary of our results is discussed in Section 7

2 Two Simple Examples

We begin by considering a pair of simple testing problems: two sided hypothesis tests for the sample mean and sample median. To illustrate the basic approach, in the case of the sample mean a particularly stark contrast is considered, where a cluster-robust variance estimator is used when observations are in fact i.i.d. We then apply the same procedure to the problem of choosing the kernel density estimator when conducting inference on the median.

2.1 Cluster-robust inference

Consider a sample $\{X_{gi}\}$, where i denotes observation i in group g. There are G groups, each containing m observations, for a total of $Gm = n$ observations.[1] We know that the observations are independent across groups and within groups, but a worried researcher suggests that we should use cluster-robust methods. For all g, i, we have that $\mathbb{E}\, X_{gi} = \mu$, $\text{Var}(X_{gi}) = \sigma^2$. Let γ and κ denote the skewness and kurtosis respectively. We would like to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. There are two ways to construct a valid Wald test based on the sample mean:

$$\bar{X}_n := \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{m} X_{gi}$$

We compare the test statistic we prefer, the classic Wald test-statistic, to a cluster robust version. For simplicity, we do not include any degrees-of-freedom correction, which will be unimportant asymptotically. The classic Wald test statistic, assuming homskedasticity, is given by:

$$W_h = \frac{(\bar{X}_n - \mu_0)^2}{\hat{\sigma}_h^2}, \quad \hat{\sigma}_h^2 = \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{m} (X_{gi} - \bar{X}_n)^2 \tag{1}$$

When discussing the asymptotic approach taken here, degrees of freedom corrections become irrelevant asymptotically, so for notational simplicity we adopt the convention of dividing by n rather than $n - 1$ when computing the variance estimator. Similarly, in the case of a cluster-robust variance estimator, often there is a degrees of freedom correction based on the number of clusters, as proposed in Hansen (2007). Since we use the large-G asymptotics of Hansen and Lee (2019), these degrees of freedom corrections disappear in the limit. Thus, the cluster-robust Wald statistic is:

$$W_c = \frac{(\bar{X}_n - \mu_0)^2}{\hat{\sigma}_c^2}, \quad \hat{\sigma}_c^2 = \frac{1}{n} \sum_{g=1}^{G} \left(\sum_{i=1}^{m} X_{gi} \right)^2 - m\bar{X}^2 \tag{2}$$

Traditional analysis proceeds as follows. Under the null hypothesis, and without any cluster dependence, we have that:

$$nW_h \Rightarrow \chi_1^2$$
$$nW_c \Rightarrow \chi_1^2$$

This fact is a basic application of Slutsky's theorem: the numerator of each test statistic,

[1] We can also accommodate balanced designs with growing cluster sizes; this type of result is also covered in Section 4.

divided by σ^2, is asymptotically χ^2_1, and each denominator converges to σ^2 in probability. The same logic holds in the case of a sequence of local alternatives, where we consider $\mu = \mu_0 + \delta/\sqrt{n}$. In this case, Slutsky's theorem applies again: the only change is that the numerator of the test statistic is no longer correctly centered, therefore the limiting distribution is $\chi^2_1(\delta^2/\sigma^2)$.

Now, let $\mu = \Delta + \mu_0$. For $a \in \{h, c\}$, the expansion of the test statistic under a fixed alternative is:

$$W_a = \frac{(\bar{X}_n - \mu)^2}{\hat{\sigma}_a^2} + \frac{2\Delta(\bar{X}_n - \mu)}{\hat{\sigma}_a^2} + \frac{\Delta^2}{\hat{\sigma}_a^2} \tag{3}$$

The first two terms converge in probability to 0, and the last term converges to Δ^2/σ^2 in each case. Thus, one way of viewing the test statistic under a fixed alternative is as a scaled estimator of the non-centrality parameter Δ^2/σ^2. In (3), the first term on the righthand side is asymptotically negligible relative to the other two terms. Under the assumption of finite kurtosis, we can obtain a normal asymptotic distribution:

$$\sqrt{n}\left(W_a - \frac{\Delta^2}{\sigma^2}\right) \Rightarrow \mathcal{N}\left(0, \frac{\Delta^2}{\sigma^2} V_a\right) \tag{4}$$

where

$$V_h = (\kappa - 1)\frac{\Delta^2}{\sigma^2} - 4\gamma\frac{\Delta}{\sigma} + 4 \tag{5}$$

$$V_c = (\kappa - 1 + 2(n-1))\frac{\Delta^2}{\sigma^2} - 4\gamma\frac{\Delta}{\sigma} + 4 \tag{6}$$

This calculation makes the simplifying assumption that the clusters all have the same size and that size is fixed at m for all n. We will relax this assumption, but the intuition developed here will remain. Even though our observations are i.i.d., (2) involves the sum over G i.i.d. cluster-sums, whereas in (1) we sum over all n observations. There are two effects here. One is that the proper normalization for (2) is $\sqrt{G}$, rather than $\sqrt{n}$, since we are summing over G squared cluster-sums. This is because for the purposes of variance estimation, we are only using G data points. We are effectively using a fixed-fraction of our data. The other effect is that there are more terms in the sum in (2) compared with (1). When considering the probability limit, these terms have mean zero and disappear. They show up in the asymptotic variance, inflating the tails of the test statistic.

The goal of this exercise is to connect the asymptotic distribution of the test statistics to power. Let C_α be the upper α quantile of a χ^2_1 random variable. Local alternatives give a (local) asymptotic approximation to the type-II error probabilities:

$$P(nW_a < C_\alpha) \to F_{\chi^2_1(\delta^2/\sigma^2)}(C_\alpha), \quad a \in \{h, c\} \tag{7}$$

This non-central chi-square distribution is the same regardless of which variance estimator we use. Thus, the asymptotic power comparisons under local alternatives cannot distinguish between Wald tests where different consistent variance estimators are used.

One implication of (4), (5), and (6) is that under fixed alternatives the test statistics have different asymptotic distributions. It is now feasible that we can productively compare the test statistics with respect to their asymptotic relative efficiency properties. Note that $V_h < V_c$ as long as $n > 1$. We again start by consider the type-II error of the test, but we normalize the test statistic based on the asymptotic distribution in (4):

$$\begin{aligned} P(nW_a < C_\alpha) &= P\left(W_a - \frac{\Delta^2}{\sigma^2} < \frac{C_\alpha}{n} - \frac{\Delta^2}{\sigma^2}\right) \\ &= P\left(V_a^{-1/2}\sqrt{n}\left(W_a - \frac{\Delta^2}{\sigma^2}\right) < \frac{C_\alpha}{\sqrt{nV_a}} - \frac{\sqrt{n}\Delta^2}{V_a^{1/2}\sigma^2}\right) \end{aligned} \tag{8}$$

The next step is heuristic. We know that the term on the lefthand side of the inequality defining the event in (8) is asymptotically standard normal. Thus, following Hettmansperger (1972), we replace P with the standard normal CDF, Φ:

$$P\left(V_a^{-1/2}\sqrt{n}\left(W_a - \frac{\Delta^2}{\sigma^2}\right) < \frac{C_\alpha}{\sqrt{nV_a}} - \frac{\sqrt{n}\Delta^2}{V_a^{1/2}\sigma^2}\right) \hookrightarrow \Phi\left(\frac{C_\alpha}{\sqrt{nV_a}} - \frac{\sqrt{n}\Delta^2}{V_a^{1/2}\sigma^2}\right) \tag{9}$$

The step to get (9) is a heuristic that serves to motivate comparing the asymptotic variances of the test statistics as a way to compare their efficiency properties. Since the argument in (9) diverges to $-\infty$, we are not at a continuity point of the distribution. Nevertheless, this approximation is useful to us. We will use this expression to compare the relative efficiency of the Wald test using the sample variance in (1) to the Wald test using the cluster-robust variance (2); out simulation evidence in Section 5 will imply that this comparison is useful for comparing finite-sample performance. Using a well-known tail approximation result for the normal distribution (see, e.g. Nolan (2020)), we have that:

$$\lim_{n\to\infty} \frac{-\log\left(\Phi\left(\frac{C_\alpha}{\sqrt{nV_a}} - \frac{\sqrt{n}\Delta^2}{V_a^{1/2}\sigma^2}\right)\right)}{-\log\left(\Phi\left(\frac{C_\alpha}{\sqrt{nV_a}} - \frac{\sqrt{n}\Delta^2}{V_a^{1/2}\sigma^2}\right)\right)} = \lim_{n\to\infty} \frac{\left(\frac{\sqrt{n}\Delta^2}{V_h^{1/2}\sigma^2} - \frac{C_\alpha}{\sqrt{n}V_h^{1/2}}\right)^2}{\left(\frac{\sqrt{n}\Delta^2}{V_c^{1/2}\sigma^2} - \frac{C_\alpha}{\sqrt{n}V_c^{1/2}}\right)^2} \tag{10}$$

$$= \frac{V_c}{V_h} \tag{11}$$

We call this relative efficiency comparison the approximate Hodges-Lehmann relative

efficiency (AHLARE). This type of comparison originates in Hodges and Lehmann (1956) and Hettmansperger (1972). Notice two aspects from this derivation. The first is that we only end up considering the higher-order terms; the term involving C_α does not enter. We could have included a degrees-of-freedom correction there which would have not entered into the result. The second is that after discarding the lower-order terms, we only need consider the ratio of the squared arguments inside the normal distribution function. The ratio in (11) serves as a heuristic characterization of the relative rate at which type-II errors disappear for the two test statistics. Since the type-II error converges to 0, taking the negative logarithm yields terms that diverge to infinity. Looking at the expressions for V_h and V_c in (5) and (6) respectively, we notice that $V_h \leq V_c$, with equality holding only when $m = 1$, which is when the cluster-robust variance estimator simplifies to the sample variance. This implies that (11) is larger than 1. Going back to our approximation in (9), this means that we expect type-II errors to disappear more quickly when using the sample variance, relative to using the cluster-robust variance.

2.2 Inference for the median

We now turn to inference on the sample median. In this example we see a case where the variance estimator converges at a slower rate than the point estimate. Let our observations X_i have density f. We are interested in a two sided test for the median,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Under standard regularity conditions on f, the sample median $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}\left(0, \frac{1}{4f(\theta)^2}\right)$$

To construct a Wald statistic, we must estimate $f(\theta)$. A commonly used option is a kernel density estimator $\hat{f}(\hat{\theta})$, using a kernel K and a bandwidth h_n. Using this estimator, our test statistic is given by:

$$W_n = 4(\hat{\theta} - \theta_0)^2 \hat{f}(\hat{\theta})^2$$

where $\hat{\theta}$ is any sample median. Under the null hypothesis, $nW_n \Rightarrow \chi^2_1$, so asymptotically valid inference can be performed by comparing nW_n to the $1-\alpha$ quantile of a χ^2_1 distribution.

One way to assess the power properties of this test is to consider a sequence of alternatives. Consider the case that $\theta_n = \theta_0 + \delta/\sqrt{n}$, where we introduce the n index to denote that we now consider a sequence of underlying distributions. This choice is exactly the right choice

such that the test statistic is asymptotically non-central:

$$nW_n \Rightarrow \chi_1^2\left(4\delta^2 f(\theta_0)^2\right) \tag{12}$$

Under a fixed alternative $\theta_n = \theta_0 + \Delta$, we expand the test statistic in an analogous manner to (3):

$$W_n = 4(\hat{\theta} - \theta)^2 \hat{f}(\hat{\theta})^2 + 8\Delta(\hat{\theta} - \theta)\hat{f}(\hat{\theta})^2 + 4\Delta^2 \hat{f}(\hat{\theta})^2$$

Slutsky's theorem applies to the first two terms: the centered estimator is asymptotically normal, and the variance estimator is consistent. Thus, these two terms are $O_P(1/n)$ and $O_P(1/\sqrt{n})$ respectively, just as in (3). The last term converges in probability to $4\Delta^2 \hat{f}(\hat{\theta})^2$. Thus, if we apply this centering we end up with:

$$W_n = 4\Delta^2(\hat{f}(\hat{\theta})^2 - f(\theta)^2) + O_P(1/n) \tag{13}$$

Classical results on the asymptotic distribution results for kernel density estimators apply to this term. Applying the delta-method, under some additional smoothness conditions we have:

$$\sqrt{nh_n}\left(W_n - 4\Delta^2 f(\theta)^2 - \frac{h_n^2}{2} f''(\theta)\right) \Rightarrow \mathcal{N}\left(0,\, 64\Delta^4 f(\theta)^3 R_K\right) \tag{14}$$

where R_K is the roughness of the kernel:

$$R_K := \int K(u)^2 du$$

Since the variance estimator converges at a slower rate than our point estimate $\hat{\theta}$, the kernel density estimator dictates the rate of convergence of the test statistic under a fixed alternative. Since (14) depends on the kernel choice through R_K, this fixed asymptotic framework can provide guidance for the choice of $\hat{f}$. If only local alternatives are considered, first-order asymptotics do not distinguish between test statistics using different kernels.

The implications for asymptotic power can be derived in much the same way as in the case of the sample mean. We again approximate the probability of making a type-II error,

with an extra step to include the bias term:

$$\begin{aligned} P(nW_n < C_\alpha) &= P(W_n - 4\Delta^2 f(\theta)^2 - \frac{1}{2}h_n^2 f''(\theta) < \frac{C_\alpha}{n} - 4\Delta^2 f(\theta)^2 - \frac{1}{2}h_n^2 f''(\theta)) \\ &= P\left(\frac{\sqrt{nh_n}(W_n - 4\Delta^2 f(\theta)^2 - \frac{1}{2}h_n^2 f''(\theta))}{8\Delta^2 f(\theta)^{3/2}\sqrt{R_K}} < \frac{-\sqrt{nh_n}4\Delta^2 f(\theta)^2}{8\Delta^2 f(\theta)^{3/2}\sqrt{R_K}} + o(1)\right) \\ &\hookrightarrow \Phi\left(\frac{-\sqrt{nh_n}4\Delta^2 f(\theta)^2}{8\Delta^2 f(\theta)^{3/2}\sqrt{R_K}} + o(1)\right) \end{aligned} \tag{15}$$

Recall the tail approximation for the normal distribution we used in (10). Just as in that case, when considering two different variance estimators with potentially different bandwidths, say $h_{1,n}$ and $h_{2,n}$ and different kernels K_1 and K_2, we only need consider the ratio of the (higher-order) arguments entering the normal distribution function. That leaves us with the limit of the ratio for our AHLARE comparison:

$$\lim_{n\to\infty} \frac{h_{1n}R_{K_2}}{h_{2n}R_{K_1}}$$

Most conventional bandwidth selection rules only depend on the kernel through a $R_K^{1/5}$ term, and therefore the goal becomes minimizing the roughness of the kernel. This formalizes and provides justification for using smooth kernels in estimating the asymptotic variance for the median. This matches with intuition from the kernel density estimation literature that the Gaussian and Epanechnikov kernels are more efficient than the Uniform kernel.

3 Fixed-alternatives asymptotics

The previous section motivates the following derivation of the asymptotic distribution of test statistics under fixed alternatives. In this section we introduce the general setup for deriving the asymptotic distribution of test statistics under fixed alternatives in the particular case of GMM estimators. We first present a result where the asymptotic distribution of the test statistic is normal, and then discuss two modifications to the basic theory.

3.1 General case: GMM

Consider the case of efficient GMM estimators. We have a set of q moment conditions

$$\mathbb{E}\, g(X_i, \beta) = 0 \tag{16}$$

where $\beta \in \mathbb{R}^p$, $X_i \in \mathcal{X}$, $g : \mathcal{X} \times \mathbb{R}^p \to \mathbb{R}^q$, $q \geq p$. The parameter of interest is the linear

functional $\theta = \ell'\beta$ for a fixed $\ell \in \mathbb{R}^p$. The vector β is estimated via efficient GMM from an i.i.d. sample of size n. Under standard regularity condition, such as those in Newey and McFadden (1994), we have that:

$$\sqrt{n}(\hat{\beta} - \beta) \Rightarrow \mathcal{N}(0, V) \tag{17}$$

where $V = (Q'\Omega^{-1}Q)^{-1}$, $\Omega = \mathbb{E}\, g(X_i, \beta)g(X_i, \beta)'$, and $Q = \mathbb{E}\, \partial_\beta g(X_i, \beta)$. We are interested in testing the two-sided hypothesis:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0 \tag{18}$$

To form a Wald test statistic, we need to estimate Q and Ω. Consistent plug-in estimators of Ω and Q are typically used to construct an estimate of V:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(X_i, \hat{\beta})g(X_i, \hat{\beta})', \quad \hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta'} g(X_i, \hat{\beta}), \quad \hat{V} = (\hat{Q}'\hat{\Omega}^{-1}\hat{Q})^{-1} \tag{19}$$

It is then straightforward to form Wald test statistics to test (18):

$$nW_n = \frac{n(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{V}\ell} \tag{20}$$

Under the null hypothesis, under common regularity conditions we have that $nW_n \Rightarrow \chi^2_1$. Similarly, under a sequence of local alternatives $\theta_n = \theta_0 + \Delta/\sqrt{n}$, we have that:

$$nW_n \Rightarrow \chi^2_1(\Delta^2/\ell' V \ell) \tag{21}$$

a non-central chi-square distribution; under the null and local alternatives the test statistics converge to their limit at the rate n.

To derive the distribution of W_n under a fixed-alternative $\theta = \theta_0 + \Delta$, we will need to make stronger assumptions than those necessary for (17).

Assumption 3.1. *There is a unique β^* such that $\beta = \beta^*$ a satisfies (16), $\frac{1}{n}\sum_{i=1}^{n} g(X_i, \hat{\beta}) = o_P(1/\sqrt{n})$, and (17) holds for $\beta = \beta^*$.*

Rather than restating standard regularity conditions, we will assume that we are in an environment in which valid asymptotic inference can be conducted. For lower-level conditions,

see Newey and McFadden (1994) and van der Vaart (1998).

Assumption 3.2. *There exists a neighborhood $\mathcal{N}$ containing β^* such that g is twice continuously differentiable on $\mathcal{N}$, and for all l, k,*

$$\mathbb{E}\left[\sup_{\beta\in\mathcal{N}}\left\|\frac{\partial^2}{\partial\beta_l\partial\beta_k}g(X_i,\beta)\right\|\right] < \infty$$

To establish stochastic equicontinuity of the GMM objective function, typically a bounded first derivative is required. Here, we require a locally bounded second derivative, since we require the asymptotic normality of linear functionals of $\hat{Q}$. When establishing a valid (stochastic) Taylor expansion of the test statistic, we need sufficient smoothness in $\partial_\beta g(X_i,\beta)$ near β^*. The requirement of differentiability eliminates quantile regression, but we will later relax this assumption for that case.

Assumption 3.3. $0 < \mathbb{E}\,\|g(X_i,\beta)\|^4 < \infty$, *and for all* $\beta \in \mathcal{N}$, $0 < \mathbb{E}\,\|\frac{\partial}{\partial\beta}g(X_i,\beta)\|^2 < \infty$.

This assumption is almost minimal for asymptotic normality of the variance estimator. We later discuss how to characterize the limiting behavior of test statistics when g has fewer than four moments, and when $\partial_\beta g$ has fewer than two moments.

To obtain the asymptotic distribution, it is helpful to consider a generalization of (3):

$$W_n - \frac{\Delta^2}{\ell' V \ell} = \frac{(\hat{\theta}-\theta)^2}{\ell'\hat{V}\ell} \tag{22}$$

$$+\frac{2\Delta(\hat{\theta}-\theta)}{\ell'\hat{V}\ell} \tag{23}$$

$$+\frac{\Delta^2}{\ell'\hat{V}\ell} - \frac{\Delta^2}{\ell' V \ell} \tag{24}$$

The righthand side of (22) is $O_P(1/n)$, as rescaled by n this term will be asymptotically χ^2_1. Under asymptotic normality of $\hat{\theta}$, standard regularity conditions will imply that (23) will be asymptotically normal. We strengthen the original assumptions to ensure asymptotic

normality of (24). Expanding this term, we get two components depending on $\hat{\Omega}$ and $\hat{Q}$:

$$\frac{\Delta^2}{\ell'\hat{V}\ell} - \frac{\Delta^2}{\ell'V\ell} = -\frac{\Delta^2}{(\ell'V\ell)^2}\operatorname{tr}\left[\Omega^{-1}QV\ell\ell'VQ'\Omega^{-1}(\hat{\Omega}-\Omega)\right] + \frac{2\Delta^2}{(\ell'V\ell)^2}\operatorname{tr}\left[V\ell\ell'VQ'\Omega^{-1}(\hat{Q}-Q)\right] + o_P(1/\sqrt{n}) \tag{25}$$

It turns out that under our assumptions each estimator $\hat{\theta}$, $\hat{\Omega}$, and $\hat{Q}$ is asymptotically linear. For $\hat{\theta}$, this is a standard result. For the variance estimator components, examining (19) shows that the variance estimators are also sums. From Assumption 3.2, we can use a Taylor expansion and replace the estimated parameter $\hat{\beta}$ by β in each estimator in (19) and include an additional term that depends on $\hat{\beta}-\beta$ when deriving the asymptotic distribution. We can show that the test statistic is also asymptotically linear under a fixed alternative, and has a normal limit.

Theorem 1. *Under Assumptions 3.1-3.3, there exists a vector c and a positive definite matrix Ξ such that:*

$$\sqrt{n}\left(W_n - \frac{\Delta^2}{\ell'V\ell}\right) \Rightarrow \mathcal{N}\left(0,\, c'\Xi c\right) \tag{26}$$

The form of c and Ξ are given in the appendix, as their expressions are rather long. We have suppressed the dependence here, but both terms depend on the alternative Δ. This emphasizes that under fixed-alternatives, the mean and variance of the test statistic will be related. Intuitively, $c'\Xi c$ corresponds to the variance of particular linear functionals of

$$a_1'g(X_i,\beta),\; a_2'g(X_i,\beta)g(X_i,\beta)'a_2,\; a_3'\frac{\partial}{\partial\beta}g(X_i,\beta)a_4 \tag{27}$$

for constant vectors $a_1, a_2, a_3 \in \mathbb{R}^q$, $a_4 \in \mathbb{R}^p$. When considering the asymptotic distribution of the test statistic under the null hypothesis, or under fixed alternatives, asymptotically all randomness in the test statistic is coming from a normalized sum $n^{-1/2}\sum_i g(X_i,\beta)$. Now, we end up with quadratic terms $(a_2'g(X_i,\beta))^2$ and bilinear terms $a_3'\partial_\beta g(X_i,\beta)a_4$ impacting the asymptotic distribution. It is helpful to understand the effect of these terms by considering when Theorem 1 does not hold. We highlight two examples that we will discuss in more detail. The first is where $(a_2'g(X_i,\beta))^2$ does not have a second moment. In this setting, asymptotic normality will no longer hold, but in particular settings we will still be able to characterize the asymptotic distribution. The second extension is when $g(X_i,\beta)$ is non-differentiable. Following classical asymptotic theory, in some cases $\mathbb{E}\,g(X_i,\beta)$ might still be

sufficiently smooth in β. Differentiating after smoothing leads to an asymptotically valid expansion, where a non-differentiable function is approximated by a smooth function. This setting is exactly the setting previously considered for the sample median, and there a result of this smoothing was the introduction of a new infinite-dimensional nuisance parameter, the density.

We can relate (26) to an AHLARE measure just as we did in Section 2. Following that example, we provide a heuristic approximate for the log probability of a type-II error. Again, letting C_α denote the $1-\alpha$ quantile of a χ^2_1 random variable,

$$\begin{aligned} P(nW_n < C_\alpha) &= P\left(\sqrt{\frac{n}{c'\Xi c}}\left(W_n - \frac{\Delta^2}{\ell' V \ell}\right) < \frac{C_\alpha}{\sqrt{n(c'\Xi c)}} - \sqrt{\frac{n}{c'\Xi c}}\frac{\Delta^2}{\ell' V \ell}\right) \\ &\hookrightarrow \Phi\left(\frac{C_\alpha}{\sqrt{n(c'\Xi c)}} - \sqrt{\frac{n}{c'\Xi c}}\frac{\Delta^2}{\ell' V \ell}\right) \end{aligned} \tag{28}$$

We use the same tail approximation. Instead of comparing to another test-statistic, we find the right normalizing constant such that the approximate log-probabilities have a limit. This constant turns out to be $1/n$:

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n}\log\left(1 - \Phi\left(\frac{\sqrt{n}\Delta^2}{(c'\Xi c)^{1/2}\ell' V \ell} - O\left(\frac{1}{\sqrt{n}}\right)\right)\right) &= \lim_{n\to\infty}\frac{1}{n}\left(\frac{\sqrt{n}\Delta^2}{(c'\Xi c)^{1/2}\ell' V \ell} - O\left(\frac{1}{\sqrt{n}}\right)\right)^2 \\ &= \frac{\Delta^4}{(c'\Xi c)(\ell' V \ell)^2} \end{aligned} \tag{29}$$

This type of limit calculation formalizes two important facts. First, the asymptotic variance of the test-statistic in (26) plays a role in the approximate convergence rate of type-II errors to 0. Second, in the normal asymptotic theory, the rate $1/n$ implies that in our approximation type-II errors disappear at an exponential rate in n. When a large-deviation theorem holds this statement can be made precise. For examples, see Dembo and Zeitouni (2009). These kinds of results generally require existence of the moment-generating function of the random variables. Our results are approximate, but do not require this generally stronger requirement, which would preclude, for example, our later discussion of heavy-tailed where fewer than four moments exist.

We will also later consider extensions to cluster dependence. In the general case, this ends up corresponding to i.n.i.d. data. We could also consider other dependence structures, such as temporal dependence. The extension is straightforward under correct specification, however under model misspecification, the situation becomes significantly more challenging. We do not pursue temporal dependence further in this paper.

3.2 Extension to heavy-tailed data

Consider the simplest linear instrumental variables model:

$$y_i = x_i\theta + \varepsilon_i$$
$$x_i = z_i\pi + \upsilon_i$$

where $y_i,\ x_i,\ z_i \in \mathbb{R}$. The conditional moment condition is satisfied so that $\mathbb{E}((\varepsilon_i,\ \upsilon_i)'|z_i) = 0$. We will also assume strong identification: $\pi^2 \geq C > 0$. We can accommodate additional control variables by regressing y_i, x_i, and z_i on the controls, then proceeding by residual regression. With this in mind, we also assume $\mathbb{E}\, z_i = 0$. We conjecture that our results in this section extend to the case when x_i and z_i are possibly vector-valued and the model is possibly over-identified, however we leave such results to future work.

The two most popular tests of the hypothesis

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$

are the Wald test, using the 2SLS estimator $\hat{\theta}$, and the Anderson-Rubin test. We first consider the Wald test. Let nW_n be the heteroskedastic-robust Wald test statistic defined as:

$$nW_n = \frac{n(\hat{\theta} - \theta_0)^2 \left(\frac{1}{n}\sum_{i=1}^n z_i x_i\right)^2}{\frac{1}{n}\sum_{i=1}^n z_i^2(y_i - x_i\hat{\theta})^2} \tag{30}$$

For the purposes of this section, our goal is to replace Assumption 3.3, while preserving validity of the Wald test. To do this, we will also replace Assumption 3.1, which can be concisely stated in the special case of a linear model:

Assumption 3.4. *There exists an $\eta > 0$ such that:*

$$\mathbb{E}\,|z_i|^{2+\eta} < \infty$$
$$\mathbb{E}\,|\varepsilon_i|^{2+\frac{4}{\eta}} < \infty$$
$$\mathbb{E}\,|\upsilon_i|^{1+\frac{2}{\eta}} < \infty$$

Further, assume that there exists a $C > 0$ such that $\pi^2 \geq C$.

For the purposes of this paper we assume strong identification. The moments conditions

in Assumption 3.4 convey that once we consider the possibility that fourth order moments may not exist, there are tradeoffs in how heavy we can allow the tails of z_i, ε_i and υ_i to be. For example, if z_i is bounded, then ε_i only needs to have a finite variance, and υ_i only needs to be integrable for these conditions to be satisfied. When $\eta = 2$, then we need at least a finite fourth moment for ε_i and a finite variance for υ_i. Under these conditions, the 2SLS estimator is asymptotically normal with the standard asymptotic variance, and the standard heteroskedastic-robust variance estimator is consistent. Placing this environment in the setup of Section 3.1, $g(X_i, \beta) = z_i(y_i - x_i\theta)$, $\Omega = \mathbb{E}\, z_i^2\varepsilon_i^2$, and $Q = \mathbb{E}\, z_i x_i$.

For a useful asymptotic distribution theory to hold when insufficient moments exist for normal limits, we consider generalizations of the central limit theorem. An important property for random variables to have in this setting is regularly varying tails. We say that X has regularly varying tails with tail index γ if for $x > 0$,

$$\lim_{t\to\infty} \frac{P(|X| > tx)}{P(|X| > t)} = x^{\gamma} \tag{31}$$

This condition can be interpreted as saying that the distribution of X is approximately polynomial far out in the tail, up to a function that is changing much more slowly than a polynomial. It turns out that (31) is the right condition for heavy-tailed random variables to still satisfy a central limit theorem, see Durrett (2019) and Nolan (2020). When (31) is satisfied, we say X is in the domain of attraction of a γ-stable law. The γ-stable laws are exactly the set of possible limit distributions for the normalized sums of random variables. The stable laws include the normal distribution and Cauchy distribution as special cases.

For our purposes, we will focus on domains of attraction under a stronger condition:

$$\lim_{t\to\infty} t^{\gamma} P(|X| > t) = c \tag{32}$$

for some constant $c > 0$. We will call this condition the approximate Pareto condition, as it implies that approximating the tails of $|X|$ with a Pareto distribution becomes arbitrarily accurate as we go farther out into the tails. Random variables satisfying (32) are said to be in the domain of normal attraction of a γ-stable law.[2]

Of particular interest to us are the completely right-skewed stable laws, which we denote S_γ. These random variables have approximately polynomial right tails. In this section we essentially consider two cases. The first is when z_i or ε_i have heavy tails, and the second is when υ_i is heavy-tailed. We begin with the first case:

[2]The terminology can be slightly confusing, as the normal distribution corresponds to $\gamma = 2$, but this is the standard phrase. See Nolan (2020) for a discussion.

Assumption 3.5. *$z_i^2\varepsilon_i^2$ is in the domain of normal attraction of a γ-stable law for $\gamma \in (1,2)$. Further, we have the relative tail condition:*

$$\lim_{t\to\infty} \frac{P(|z_i x_i| > t)}{P(z_i^2\varepsilon_i^2 > t)} = 0 \tag{33}$$

Assumption 3.5 fulfills the same purpose here as Assumption 3.3; both are sufficient for a central limit theorem to hold. In this case, Assumption 3.5, specifically (32) ensures that there exists a constant b such that:

$$bn^{-1/\gamma} \sum_{i=1}^{n} (z_i^2\varepsilon_i^2 - \Omega) \Rightarrow S_\gamma \tag{34}$$

Assumption 3.5 also places some restrictions on η in Assumption 3.4, specifically requiring $\eta < 2$. If $\eta \geq 2$, then $z_i^2\varepsilon_i^2$ would be in the domain of attraction of a normal distribution, so that $b_n = n^{-1/2}$ and $\gamma = 2$.

The second part of Assumption 3.5,(33), rules out cases where $z_i x_i$ has very heavy tails relative to $z_i^2\varepsilon_i^2$. Since $x_i = z_i\pi + \upsilon_i$, this rules out cases where υ_i has very heavy tails, or the tails of ε_i are sufficiently thin such that $z_i^2\varepsilon_i^2$ and z_i^2 have similar tail properties. This condition is satisfied when ε_i is conditionally (on z_i) heteroskedastic and the conditional heteroskedasticity function is non-decreasing in $|z_i|$. In general, if z_i and ε_i are independent, then $z_i^2\varepsilon_i^2$ and z_i^2 will have regularly varying tails of the same order; such a result is due to Breiman (1965). If there is dependence, the tails of the product $z_i^2\varepsilon_i^2$ will be heavier than those of z_i^2. See Resnick (2007) for more details. Since most empirical work in econometrics assumes that conditional heteroskedasticity is present, these appear to be relevant cases.

An insight that we believe is new to this paper is that heavy-tailed instruments or regressors can have serious consequences for the power of tests. Conventional wisdom is that thick-tailed regressors lead to more-precise estimates of the parameter of interest. This intuition comes from the homoskedastic regression model. Recent work considering heavy-tails in regression has focused on the quality of the normal approximation. In Shephard (2020) an alternative procedure is proposed where a bounded transformation of the regressors is used. Müller (2020) proposes an alternative testing procedure also designed to improve size control. When heteroskedasticity is present, heavy-tailed regressors can also negatively impact power.

Heavy-tailed regressors also provide a model for high-leverage designs. The leverage values in this model are $h_{ii} = z_i^2 / \sum_{i=1}^n z_i^2$. It is well-known that when instruments are high-dimensional, or reflect a sparse-dummmy variable structure, that high leverage values can

correspond to poor inference; see Anatolyev (2019) for a review. Recently, Young (2021) has argued that in many cases of empirical interest, high (sample) kurtosis of the instruments in 2SLS is a highly correlated of high leverage. The connection with heavy-tailed data is that when z_i^2 is in the domain of attraction of a stable law, $\max_i h_{ii}$, after centering, will converge in distribution to a random variable. This result, due to Chow and Teugels (1978), suggests that when high leverage values and heteroskedasticity are present, there can be implications for power. We summarize those implications in the following theorem.

Theorem 2. *Under Assumptions 3.4 and 3.5 there exists a constant b, such that:*

$$\frac{\Omega^2 b}{\Delta^2 Q^2} n^{1-1/\gamma} \left(W_n - \frac{\Delta^2 Q^2}{\Omega} \right) \Rightarrow -S_\gamma \tag{35}$$

Following our plan laid out in Section 2, we relate the asymptotic distribution of simple test statistics to AHLARE:

$$\begin{aligned} P(nW_n < C_\alpha) &= P\left(bn^{1-1/\gamma} \left(W_n - \frac{\Delta^2 Q^2}{\Omega^2} \right) < \frac{bC_\alpha}{n^{1/\gamma}} - \frac{bn^{1-1/\gamma}\Delta^2 Q^2}{\Omega} \right) \\ &\hookrightarrow 1 - F_\gamma \left(bn^{1-1/\gamma}\Omega - \frac{bC_\alpha}{n^{1/\gamma}} \right) \end{aligned}$$

where F_γ is the distribution function of a S_γ random variable. Similar to the case when the limit was normal, we can asymptotically approximate the tail probabilities of a stable law:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{\log n} \log &\left(1 - F_\gamma \left(b_n n^{1-1/\gamma}\Omega - \frac{b_n C_\alpha}{n^{1/\gamma}} \right) \right) \\ &= \lim_{n\to\infty} \frac{1}{\log n} \log \left(n^{\gamma-1} \right) \\ &= \gamma - 1 \end{aligned} \tag{36}$$

For details on the approximation in (36), see Nolan (2020). In comparing this result with (29), note the significantly slower rate of convergence: $\log n$. Thus, the approximate type-II errors disappear linearly in n when the tails of $z_i^2 \varepsilon_i^2$ are sufficiently heavy.

An important implication of this result is that when $\mathbb{E}(\varepsilon_i | z_i) = 0$, rather than using z_i as an instrument, there could be benefits of using a bounded transformation of z_i. This idea is explored in Shephard (2020), and here we see how this could also improve power in some

cases. Let ψ be such a transformation, so that our moment condition is now:

$$\psi(z_i)(y_i - x_i\theta)$$

Examples of ψ include the sign function, as in Shephard (2020), but the intuition follows for any bounded transformation, such as the Huber-score, the normal CDF, or the sigmoid function. For asymptotic normality of the test statistic under fixed alternatives, with a bounded instrument we only require $\mathbb{E}\,\varepsilon_i^4 < \infty$ and $\mathbb{E}\,\upsilon_i^2 < \infty$. Thus, there will be cases in which using $\psi(z_i)$ instead of z_i will lead to approximately exponential rates of convergence of type-II error probabilities to zero rather than the linear rates implied by (36).

An important example of this phenomenon is the case of linear regression, i.e. $x_i = z_i$. Typical intuition from linear regression implies that higher-variance regressors lead to more accurate inference. Re-scaling the regressors by a constant $C > 1$ leads to a decrease in the asymptotic variance by C^{-1}. Under local asymptotics, this corresponds to an increase in local power. Our results imply that when considering asymptotic power under heteroskedasticity the benefit of highly variable x_i becomes less clear. As this is the relevant case in much of econometric research, our result indicates that situations with heavy-tailed regressors (or instruments) should be treated with care.

We now consider the case where υ_i has heavy tails. Here, we will compare the Wald test to the Anderson-Rubin test. Through the rest of this section, we will assume homoskedasticity and use the homoskedastic version of each test statistic. We define $\sigma^2 = \mathbb{E}\,\varepsilon_i^2$, $\sigma_\upsilon^2 = \mathbb{E}\,\upsilon_i^2$, and $\rho = \mathbb{E}\,\varepsilon\upsilon$. Let $\hat{\pi}$ be the least-squares estimator of π from the first-stage regression. Then, the test statistics we consider are:

$$nW_n^h := \frac{n(\hat{\theta} - \theta_0)^2\hat{Q}^2}{\frac{1}{n}\sum_{i=1}^n (y_i - x_i\hat{\theta})^2}, \qquad nR_n^h := \frac{n(\hat{\theta} - \theta_0)^2\hat{Q}^2}{\frac{1}{n}\sum_{i=1}^n ((y_i - z_i\hat{\pi}\hat{\theta}) - (x_i - z_i\hat{\pi})\theta_0)^2} \tag{37}$$

where in each case we reject when the test statistic is larger than C_α. We note that the Anderson-Rubin test statistic is not a Wald test, but we can interpret the denominator of R_n^h as a variance estimator which is consistent under the null hypothesis, so we can still easily adapt our discussion to this comparison as well. The Anderson-Rubin test is designed to be robust to weak instrument cases where $\pi \approx 0$. In Staiger and Stock (1994), attention is focused on the concentration parameter. We define the concentration parameter here as:

$$\mu^2 := \frac{\pi^2\,\mathbb{E}\,z_i^2}{\mathbb{E}\,\upsilon_i^2} \tag{38}$$

We replace our definition of the concentration parameter is slightly different from the

the expression in Staiger and Stock (1994), but it can be viewed as an unconditional version of their concentration parameter. The traditional weak instrument literature considers the case where μ^2 is small because π^2 is small. When our first-stage errors have heavy tails, (38) may be small even when π^2 is large, or even 0 when $\mathbb{E}\, v_i^2 = \infty$, which is not precluded by Assumption 3.4. In particular, if z_i is binary 0-1 with $\mathbb{E}\, z_i = \eta$, then $\mu^2 = \pi^2\eta(1-\eta)/\mathbb{E}\, v_i^2$, and η in Assumption 3.4 can be arbitrarily large, so that $\mu^2 = 0$ can occur while inference with 2SLS or Anderson Rubin is still valid. We will provide a theorem that implies the Anderson-Rubin test is not robust to heavy-tailed first-stage errors relative to the Wald test. This behavior is not captured when using local-asymptotics, as both test statistics, when $\pi^2 > 0$, have the same asymptotic distribution under the null hypothesis and local alternatives of the form $\theta_n = \theta_0 + \delta/\sqrt{n}$.

To understand why Anderson-Rubin tests are more sensitive to heavy-tails in the first-stage, consider the expansion of the squared residuals in the variance estimator in the denominator of nR_n^h in (37) when $\theta = \theta_0 + \Delta$:

$$(y_i - z_i\hat{\pi}\hat{\theta}) - (x_i - z_i\hat{\pi})\theta_0 = \varepsilon_i + \Delta v_i + z_i(\hat{\pi} - \pi)\Delta - z_i\hat{\pi}(\hat{\theta} - \theta)$$

Thus, the asymptotic distribution of R_n will involve a term governed by v_i^2, since $\Delta \neq 0$. Now, for comparison, the residuals in the denominator of nW_n in (37) are:

$$y_i - x_i\hat{\theta} = \varepsilon_i - z_i\pi(\hat{\theta} - \theta) - v_i(\hat{\theta} - \theta)$$

Here, consistency of the 2SLS estimator will imply that we will only need some weak moment existence conditions to be met.

Assumption 3.6. *We assume that v_i^2 is in the domain of normal attraction of a γ-stable law with $\gamma \in (1,2)$, and $\mathbb{E}\, \varepsilon_i^4 < \infty$.*

Note that this will imply that $\mathbb{E}\,|v_i|^{2\zeta} < \infty$, for $\zeta \in (1,\gamma)$. It also implies that $\mathbb{E}\,|v_i|^{2+\eta'} = \infty$ for some $\eta' \geq 2(\gamma - 1)$. We are now ready to state a theorem:

Theorem 3. *Under Assumption 3.4 and 3.6, we have that there exists $V, a > 0$ such that:*

$$\sqrt{n}\left(W_n - \frac{\Delta^2 Q^2}{\sigma^2}\right) \Rightarrow \mathcal{N}(0, V) \tag{39}$$

$$an^{1-1/\gamma}\left(R_n - \frac{\Delta^2 Q^2}{\sigma^2 + 2\Delta\rho + \Delta^2\sigma_v^2}\right) \Rightarrow -S_\gamma \tag{40}$$

For comparing the two test statistics in this setting we will see that the constants are not important. Our discussion of this result mirrors our previous discussion when the limit distribution of a test statistic was stable. Consider that the convergence in (39) is a special case of (26). Thus, analogously to the derivations leading to (29), we have that, for large n,

$$-\log P(nW_n^h < C_\alpha) \approx \frac{n\Delta^4 Q^4}{V\sigma^4} \tag{41}$$

For comparison, the rate in (40) follows the same heavy-tailed case in (36). Thus, our heuristic for the log probabilities, for large n, implies

$$-\log P(nR_n^h < C_\alpha) \approx (\gamma - 1)\log(n) \tag{42}$$

Thus, we expect the Anderson-Rubin test to perform quite poorly relative to Wald tests when the first stage has heavy-tailed errors. Combining (41) and (42), we see that:

$$\lim_{n\to\infty} \frac{\log P(nR_n^h < C_\alpha)}{\log P(nW_n^h < C_\alpha)} \approx \lim_{n\to\infty} \frac{(\gamma - 1)\log(n)}{(n\Delta^4 Q^4)/(V\sigma^4)} = 0 \tag{43}$$

Thus, when the first-stage errors are sufficiently heavy-tailed and identification is strong, we expect a stark difference in the asymptotic power properties of Wald tests compared with Anderson-Rubin tests.

3.3 Extension to quantile regression

In this section we apply our procedure to variance estimation in the context of quantile regression. Note that quantile regression corresponds to GMM with moment conditions:

$$g(y_i, x_i, \beta_\tau) = x_i\left(\tau - \mathbb{1}_{[y_i \leq x_i'\beta_\tau]}\right)$$

This function is not continuous, much less differentiable. The smoothing procedure to obtain a replacement for the matrix of partial derivatives Q introduces an infinite dimensional nuisance parameter which is present during variance estimation, but does not play a role in

estimating β_τ. In this section, we assume $\{(y_i,\, x_i')\}_{i=1}^n$ are i.i.d.. The model we work with is:

$$y_i = x_i'\beta_\tau + \varepsilon_i, \quad Q_\varepsilon(\tau|x_i) = 0 \tag{44}$$

where $Q_\varepsilon(\cdot|x_i)$ is the conditional quantile function of ε_i. Let $f(\cdot|x_i)$ be the conditional density of ε_i given x_i. We define:

$$\Omega_\tau := \tau(1-\tau)\,\mathbb{E}\, x_i x_i'$$
$$Q_\tau := \mathbb{E}[f(0|x_i)x_i x_i']$$

Assumption 3.7. *Suppose that for an estimator $\hat{\beta}_\tau$:*

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \Rightarrow \mathcal{N}\left(0,\, Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\right)$$

Estimating Ω_τ is straightforward: under correct specification, we set:

$$\hat{\Omega}_\tau = \frac{1}{n}\sum_{i=1}^n \tau(1-\tau)x_i x_i' \tag{45}$$

Assuming $\mathbb{E}\,\|x_i\|^4 < \infty$, for any constant matrix C, $\sqrt{n}\,\mathrm{tr}(C(\hat{\Omega}_\tau - \Omega_\tau)) = O_P(1)$ by a standard central limit theorem. We will see that in our setting, the distribution of test statistics will only depend on properties of $\hat{Q}_\tau$, not $\hat{\Omega}_\tau$.[3] We will need to make several assumptions on the kernel used, the conditional density of ε_i, and the bandwidth choice h_n.

Assumption 3.8. *The kernel function K is symmetric, of bounded variation, and normalized such that:*

$$\int_{\mathbb{R}} uK(u)du = 0, \qquad \int_{\mathbb{R}} u^2K(u)du = 1$$

This first assumption is satisfied by all kernel functions used in practice, such as the Gaussian, Epanechnikov, Uniform, Biweight, and Triweight kernels. This assumption implies that the function can only rise and fall finitely many times.

[3]We can actually relax the conditional-quantile assumption here, as $\hat{\Omega}_\tau$ will still be $\sqrt{n}$-consistent for its probability limit so the asymptotic distribution of the test statistic is unchanged.

Assumption 3.9. *There exist functions $G_j(x_i)$ such that for all x_i, $G_j(x_i) \geq |f^{(j)}(u|x_i)|$, uniformly in u, $j \in \{0, 1, 2\}$. Furthermore, G_j also satisfy, for some $\delta_j > 0$, $\mathbb{E}(G_0(x_i)\|x_i\|^{4+\delta_0}) < \infty$, $\mathbb{E}(G_1(x_i)\|x_i\|^{2+\delta_1}$, and $\mathbb{E}(G_2(x_i)\|x_i\|^2) < \infty$.*

This assumption is quite similar to assumptions used in Kato (2012) in proving asymptotic normality of the variance estimator when using the uniform kernel. Bounding the density and the first two derivatives is standard in the literature on kernel density estimation, and in the regression context due to the conditional nature of the density we must impose additional restrictions on the regressors to ensure integrability of the envelope functions that are used in the bounds.

Assumption 3.10. $h_n = o(\log n/\sqrt{n})$.

This bandwidth condition will allow the rate-optimal bandwidth, $h_n \propto n^{-1/5}$. It is slightly stronger than the bandwidth condition in Powell (1991), $n^2 h_n \to \infty$. Consider testing a linear hypothesis of the form $H_0 : \ell'\beta_\tau = \theta_0$. Using a kernel estimator of Q_τ, the Wald statistic is nW_n, where:

$$W_n = \frac{(\ell'\hat{\beta}_\tau - \theta_0)^2}{\ell'\hat{Q}_\tau^{-1}\hat{\Omega}_\tau\hat{Q}_\tau^{-1}\ell}, \qquad \hat{Q}_\tau = \frac{1}{nh_n}\sum_{i=1}^{n} x_i x_i' K\left(\frac{\hat{\varepsilon}_i}{h_n}\right)$$

We also define the matrix A:

$$A = \frac{2Q_\tau^{-1}\Delta^2\ell\ell'Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}}{(\ell'Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell)^2}$$

Under these assumptions, we can prove the following theorem:

Theorem 4. *Under Assumptions 3.8-3.10, we have that:*

$$\sqrt{nh_n}\left(W_n - \frac{\Delta^2}{\ell'Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} - B_n\right) \Rightarrow \mathcal{N}\left(0,\, \mathbb{E}[(x_i'Ax_i)^2 f(0|x_i)R_K]\right) \tag{46}$$

where $R_K = \int K(u)^2 du$ is the roughness, and the bias term is:

$$B_n = \frac{1}{2}\,\mathbb{E}[x_i'Ax_i f''(0|x_i)h_n^2] \tag{47}$$

The proof is distinct from that in Kato (2012), in that both proofs utilize empirical process methods, but here we do not employ combinatorial arguments directly. Rather, we use kernel properties from Giné and Nickl (2016) and a maximal inequality from Chernozhukov et al. (2014). Note that for the asymptotic distribution, there is no contribution from estimating β_τ or Ω_τ. Since both of these terms converge at the $\sqrt{n}$-rate, compared with Q_τ they can effectively be treated as known, using empirical process methods to bound those types of errors.

Using (46), we use our approach to provide guidance for the choice of kernel, based on comparing AHLARE for different kernel choices. In general, the choice of kernel has seemed unimportant. When only consistency matters for variance estimation, we should choose the simplest kernel to implement. Here we provide justification for using kernels other than the uniform kernel.

Proposition 1. *Using the Epanechnikov kernel is AHLARE-optimal for estimating the asymptotic variance for the quantile regression esitimator. In addition, the Gaussian kernel is more efficient than the rectangular kernel.*

To see why this is the case, consider approximating the type-II error of this test, just as in (28), noting that $B_n = O(h_n^2)$,

$$P(nW_n < C_\alpha) = P\left(\sqrt{nh_n}\left(W_n - \frac{\Delta^2}{\ell' Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} - B_n\right) < -\frac{\sqrt{nh_n}\Delta^2}{\ell' Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} - O(n^{1/2}h_n^{5/2})\right)$$
$$\hookrightarrow 1 - \Phi\left(\frac{\sqrt{nh_n}\Delta^2}{\sqrt{\mathbb{E}[(x_i'Ax_i)^2 f(0|x_i)R_K]}\ell' Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} + O(n^{1/2}h_n^{5/2})\right)$$

Using our general formulation in (29), we can see that we want to minimize:

$$\frac{\sqrt{nh_n}\Delta^2}{\sqrt{\mathbb{E}[(x_i'Ax_i)^2 f(0|x_i)R_K]}\ell' Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} \tag{48}$$

Now, generally, the bandwidth choice depends on R_K, but generally only up to a factor of $R_K^{1/5}$. Thus, to minimize the variance of the test statistic, we want to minimize R_K, leading to the recommendation to use the Gaussian or Epanechnikov kernels in practice. In Powell (1991), consistency of the kernel variance estimator was proved for the choice of the uniform kernel. In Stata, the default kernel when using the `qreg` command is the Epanechnikov kernel, while in R, in the package `quantreg`, the default when using `rq` is the Gaussian

kernel. These choices were based on traditional intuition from the general kernel density estimation problem, but there was no theoretical reason to prefer these smooth kernels over the uniform kernel in the testing problem. We provide such a justification here for using smooth kernels in the context of estimating the asymptotic variance of the quantile regression parameter vector.

4 Extension to Cluster-dependent data

In this section we extend the results of Section 3.1 to cluster-dependent data. We focus on two empirically relevant cases: linear regression models and linear instrumental variable models. We present an extension of Theorem 1, and look at the asymptotic behavior of the test statistic when irrelevant clusters are imposed. This is empirically relevant, as we demonstrate that when a fine cluster is appropriate, using a coarser cluster will lead to asymptotic efficiency loss. We now turn our attention to ordinary least-squares (OLS) and two-stage least squares (2SLS) and consider the choice between two levels of clustering. This is a common choice empirical researchers must make. Common considerations are the classroom or school level for student-level data, and county, state, or region in common U.S. survey data.

We consider here the just-identified case for 2SLS, and for our tests we will focus on linear functionals $\theta = \ell'\beta$. The extension to over-identified settings and nonlinear restrictions is conceptually straightforward, if more notationally cumbersome. The model is:

$$y_{dgi} = x'_{dgi}\beta + \varepsilon_{dgi}, \quad \mathbb{E}[\varepsilon_{g(d)}|Z_{g(d)}] = 0$$

where $Z_{g(d)}$ and $X_{g(d)}$ will generally be the $n_{g(d)} \times p$ matrices with row i equal to z'_{dgi} or x'_{dgi} respectively, and y_{dgi} and $\varepsilon_{g(d)}$ will be $n_{g(d)}$ vectors. The case of OLS is nested with $Z_{g(d)} = X_{g(d)}$. Consider two levels of clustering: for example, classrooms versus schools. We will denote the number of students in classroom g by $n_{g(d)}$, and the number of students in school d by $n_{\bullet d} = \sum_{g=1}^{G_d} n_{g(d)}$, where G_d denotes the number of classrooms in school d. The total number of observations is $n = \sum_{d=1}^{D}\sum_{g=1}^{G_d} n_{g(d)} = \sum_{d=1}^{D} n_{\bullet d}$. The truth is that observations are independent across classrooms, but this is unknown to the researcher. We

define:

$$\Omega_n := \frac{1}{n}\sum_{d=1}^{D}\sum_{g=1}^{G_d}\mathbb{E}(Z'_{g(d)}\varepsilon_{g(d)}\varepsilon'_{g(d)}Z_{g(d)})$$
$$Q_n := \frac{1}{n}\sum_{d=1}^{D}\sum_{g=1}^{G_d}\mathbb{E}(Z'_{g(d)}X_{g(d)})$$
$$V_n := Q_n^{-1}\Omega_n(Q'_n)^{-1}$$

We will need to make some assumptions to obtain not only validity of the Wald test statistic, but asymptotic normality under fixed alternatives.

Assumption 4.1. *For some $2 \leq r_A < \infty$, $A \in \{G, D\}$, there exist C_G, C_D such that:*

$$\frac{\left(\sum_{d=1}^{D}\sum_{g=1}^{G_d} n_{g(d)}^{2r_G}\right)^{2/r_G}}{n} \leq C_G < \infty, \quad \frac{\left(\sum_{d=1}^{D} n_{\bullet d}^{2r_D}\right)^{2/r_D}}{n} \leq C_D < \infty$$

$$\lim_{n\to\infty}\max_{g,d}\frac{n_{g(d)}^4}{n} = \lim_{n\to\infty}\max_{d\leq D}\frac{n_{\bullet d}^4}{n} = 0$$

This first assumption places restrictions on how quickly the clusters can grow with n and how heterogenous the clusters can be. Equal-sized clusters are allowed, as well as clusters that grow as a power of n, such as $n_{g(d)} = n^\omega$, for $\omega \in (0, 1)$. The same holds true for $n_{\bullet d}$ as well. For a more complete discussion, see Hansen and Lee (2019).

For the next assumption, we introduce some notation:

$$a_n := (Q_n^{-1})'\ell \tag{49}$$
$$b_n := V_n\ell \tag{50}$$

We then define:

$$Y_{g(d)} := \begin{pmatrix} Z_{g'(d)}\varepsilon_{g(d)} \\ a_n'(Z_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'Z_{g(d)} - \mathbb{E}\, Z_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'Z_{g(d)})a_n \\ a_n'(Z_{g(d)}'X_{g(d)} - \mathbb{E}\, Z_{g(d)}'X_{g(d)})b_n \end{pmatrix} \tag{51}$$

$$Y_{\bullet d} := \begin{pmatrix} \sum_{g=1}^{G_d} Z_{g(d)}'\varepsilon_{g(d)} \\ a_n'\left(\left[\sum_{g=1}^{G_d} Z_{g(d)}'\varepsilon_{g(d)}\right]\left[\sum_{g=1}^{G_d} Z_{g(d)}'\varepsilon_{g(d)}\right]' - \sum_{g=1}^{G_d}\mathbb{E}\, Z_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'Z_{g(d)}\right)a_n \\ a_n'\left(\sum_{g=1}^{G_d} Z_{g(d)}'X_{g(d)} - \sum_{g=1}^{G_d}\mathbb{E}\, Z_{g(d)}'X_{g(d)}\right)b_n \end{pmatrix} \tag{52}$$

$$\tag{53}$$

The main idea behind the results here is deriving central limit theorems based on sums of these mean-zero vectors. Notice that the sums $\sum_d \sum_g Y_{g(d)}$ and $\sum_d Y_{\bullet d}$ will have the same first q entries and the same last entry, but the second to last will be different between the two sums. We define:

$$\Xi_n^G := \frac{1}{n^2}\sum_{d=1}^{D}\sum_{g=1}^{G_d}\mathbb{E}\, Y_{g(d)}Y_{g(d)}' \tag{54}$$

$$\Xi_n^D := \frac{1}{n^2}\sum_{d=1}^{D}\mathbb{E}\, Y_{\bullet d}Y_{\bullet d}' \tag{55}$$

It turns out that all entries will be equal across these two matrices except for the second to last diagonal entry, which corresponds to the variance estimation. We will require that Ξ_n^G is well-behaved, and require nothing further since $\Xi_n^D - \Xi_n^G$ is positive semi-definite.

Assumption 4.2.

1. $\lambda_{min}(\Omega_n) \geq \lambda > 0$ *and* Q_n *has rank* p.
2. $\lambda_{min}(\Xi_n^G) \geq \lambda > 0$.

The first part of this assumption places some restrictions on the design, and these conditions are sufficient for identification of θ, and non-degeneracy of the asymptotic distribution. The second part is a non-degeneracy requirement for the components of the test statistic. This non-degeneracy will be satisfied in almost all cases, and seems to be a mild assumption, but it is stronger that what is required for validity of the test statistic.

Assumption 4.3. *For r_G, r_D in Assumption 4.1, there exists $\max\{r_G, r_D\} < s/2 < \infty$ such that $\sup_{i,g,d} \mathbb{E}\,|y_{dgi}|^{2s} < \infty$, $\sup_{i,g,d} \|x_{dgi}\|^{2s} < \infty$, and $\sup_{i,g,d} \mathbb{E}\,\|z_{dgi}\|^{2s} < \infty$.*

This final assumption ensures the necessary uniform integrability condition is satisfied to apply a Lindeberg central limit theorem. This assumption is quite strong; essentially, 8 moments are required to exist for the observed random variables. This is not surprising when we consider that for validity of heteroskedastic-robust inference, we generally assume fourth moments exist y_{dgi}, x_{dgi}, and z_{dgi}. In our case, we also need the variances of the squared terms to exist, which implies that we will double the number of required moments. For estimating the variance, we define two different plug-in estimators:

$$\begin{aligned}
\hat{Q}_n &= \frac{1}{n}\sum_{d=1}^{D}\sum_{g=1}^{G_d} Z_{g(d)}' X_{g(d)} \\
\hat{\Omega}_n^G &= \frac{1}{n}\sum_{d=1}^{D}\sum_{g=1}^{G_d} Z_{g(d)}' \hat{\varepsilon}_{g(d)} \hat{\varepsilon}_{g(d)}' Z_{g(d)} \\
\hat{\Omega}_n^D &= \frac{1}{n}\sum_{d=1}^{D}\left(\sum_{g=1}^{G_d} Z_{g(d)}' \hat{\varepsilon}_{g(d)}\right)\left(\sum_{g=1}^{G_d} \hat{\varepsilon}_{g(d)}' Z_{g(d)}\right) \\
\hat{V}_n^G &= \hat{Q}_n^{-1}\hat{\Omega}_n^G(\hat{Q}_n')^{-1} \\
\hat{V}_n^D &= \hat{Q}_n^{-1}\hat{\Omega}_n^D(\hat{Q}_n')^{-1}
\end{aligned}$$

We then construct the standard Wald test statistic for testing $H_0 : \ell'\beta = \theta_0$ against a two-sided alternative:

$$nW_n^G = \frac{n(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{V}_n^G\ell}$$

Before stating the theorem, we will also define:

$$\begin{aligned}
c_n &:= \frac{1}{n}\sum_{d=1}^{D}\sum_{g=1}^{G} \mathbb{E}[X_{g(d)}' Z_{g(d)} a_n a_n' Z_{g(d)}' \varepsilon_{g(d)}] \\
\xi_n &:= \Delta/\ell' V_n \ell \\
\nu_n &:= \begin{pmatrix} 2(\xi_n a_n - (Q_n')^{-1}c_n) \\ -\xi_n^2 \\ 2\xi_n^2 \end{pmatrix}
\end{aligned}$$

Our assumptions give us the following characterization of the two different test statistics under a fixed alternative $\theta = \theta_0 + \Delta$:

Theorem 5. *Under assumptions Assumptions 4.1-4.3, we have that there exist sequences* Ξ_n^A, ν_n, *such that for* $A \in \{G, D\}$,

$$\frac{1}{\sqrt{\nu_n' \Xi_n^A \nu_n}} \left(W_n^A - \frac{\Delta^2}{\ell' V_n \ell} \right) \Rightarrow \mathcal{N}(0, 1) \tag{56}$$

Furthermore, for all n, $\Xi_n^G \preceq \Xi_n^D$, *with equality if and only if* $(Q_n')^{-1}\ell = 0$.

$\nu_n' \Xi_n^G \nu_n$ is the rate associated with clustering at the finer level, e.g. classrooms, and $\nu_n' \Xi_n^D \nu_n$ is the rate associated with clustering at the coarser level, e.g. the school level. Returning to the example of the sample mean, we perform similar calculations to (11). This leads to a relative efficiency comparison based on AHLARE of the rate at which type-II errors disappear:

$$\lim_{n\to\infty} \frac{-\log P(nW_n^G < C_\alpha)}{-\log P(nW_n^D < C_\alpha)} \hookrightarrow \lim_{n\to\infty} \frac{\Xi_n^D}{\Xi_n^G} \tag{57}$$

It might appear that this comparison is ambiguous, but it turns out that the difference $\nu_n'(\Xi_n^G - \Xi_n^D)\nu_n$ has a very simple form. We first define:

$$\Pi_{g(d)} := \frac{\ell' Q_n^{-1} \frac{1}{n} \mathbb{E}[Z_{g(d)}' \varepsilon_{g(d)} \varepsilon_{g(d)}' Z_{g(d)}](Q_n')^{-1}\ell}{\ell' V_n \ell}$$

This is the proportion of total variation coming from cluster g. The difference of interest can be expressed as:

$$\nu_n'(\Xi_n^G - \Xi_n^D)\nu_n = \frac{\Delta^4}{(\ell' V_n \ell)^2} \sum_{d=1}^{D} \sum_{g=1}^{G_d} \Pi_{g(d)} \left(\sum_{h \neq g} \Pi_{h(d)} \right)$$

Thus, the penalty for over-clustering is unambiguous: $(\nu_n' \Xi_n^D \nu_n)/(\nu_n' \Xi_n^G \nu_n) > 1$ for all n. Looking at (57), this implies that the rate at which type-II errors disappear when clustering at the coarse level is slower than when we cluster at the correct level, the finer level. One scenario which leads to a larger penalty term stands out: when using the coarse clustering exacerbates underlying heterogeneity. When a particularly large Π_g is placed in a cluster with a large number of independent clusters, the effect of that cluster on the variance estimator will be inflated by a factor equal to $\sum_{h \neq g} \Pi_{h(d)}$.

It is also helpful to consider the case when the sampling scheme is i.i.d., but clusters are imposed by the researcher when estimating the asymptotic variance. In that case, each

$\Pi_{g(d)} = 1/n$, therefore we end up with:

$$\nu_n'(\Xi_n^G - \Xi_n^D)\nu_n = \frac{\Delta^4}{(\ell' V_n \ell)^2}\left[\frac{1}{n}\sum_{d=1}^{D} G_d^2 - 1\right]$$

When the cluster sizes G_d^2 are all equal, this simplifies further to the $G_d - 1$ penalty term analogous to the case of the sample mean in (6). When the cluster sizes are heterogeneous, the penalty can be much larger.

This analysis has both theoretical and practical implications. We expand upon the finite sample results in Abadie et al. (2017) by demonstrating an asymptotic penalty associated with misguided clustering. Our results, being asymptotic in nature, also hold uniformly over a broad class of data-generating processes. We also point out that our analysis answers a different counterfactual than that posed by a hypothesis test of clustering level. The test proposed in MacKinnon et al. (2020b) tests the null hypothesis that the fine clustering level is the correct level. Our analysis supposes that the fine clustering level is the correct level, and quantifies the penalty for using coarser clusters. One way to use (56) which we will discuss below is in conducting power analysis, which will allow researchers another tool for balancing robustness with efficiency in presenting their results.

We do not claim here that under-clustering is a good idea. Failing to cluster can lead to invalid inference. Our goal here is to highlight the fact that there are tradeoffs. Depending on the researcher's information regarding the sampling scheme, it would be reasonable to weigh the benefit of clustering at a coarser level (lower type-I errors) against the costs (higher type-II errors). These results formalize the costs associated with coarser "over-clustering" in this trade-off.

5 Simulation Evidence

5.1 Clustering

In this section we evaluate the finite sample predictions made by the theory we have developed up to this point. Our first setup is very simple: 1440 i.i.d observations from a $\mathcal{N}(\mu_0 + \Delta, 2)$ distribution, with cluster sizes of 72, 144, 288, 480, 720.

In Figure 1, we plot Monte-Carlo estimates of the power of a two-sided test against a null hypothesis that the mean is 0 using 10000 simulation draws. The left panel plots the power curve, and the right panel plots the percent efficiency loss relative to the standard t-test. We adjusted the critical values in these simulations so that the type-I error rate is 0.01 for all tests.

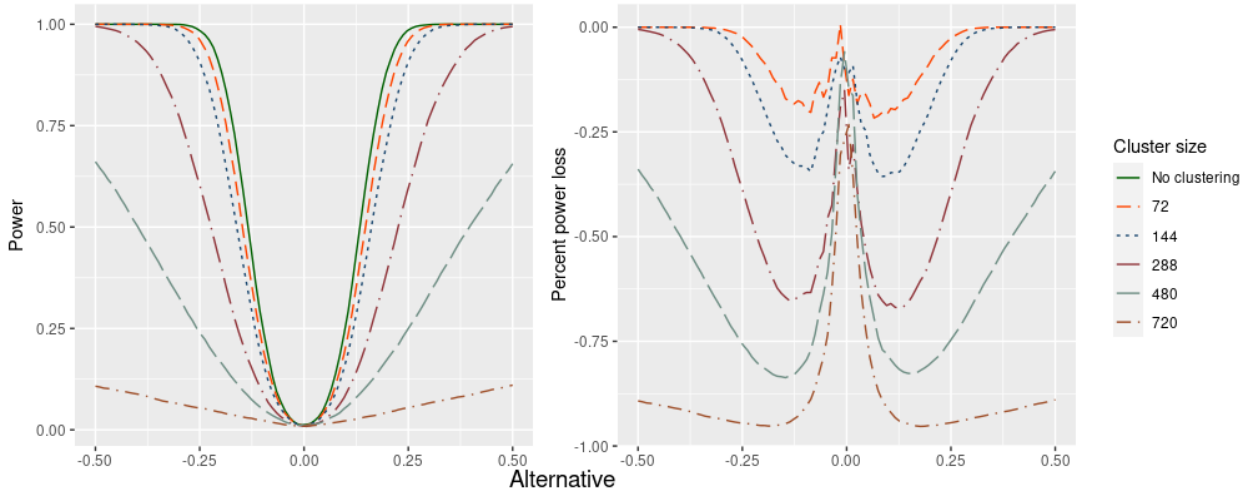

Figure 1: Sample mean, i.i.d. observations

Unsurprisingly the power of the standard t-test dominates the clustered versions. The t-test is the uniformly most powerful test, so this should be expected. Notice that the larger the cluster size is relative to the total sample size, the more the behavior seems to reflect the rate penalty for increasing cluster size rather than the fixed cost of over-clustering with a fixed cluster size. Dividing the sample into 2 or 3 clusters seems extreme, but in our empirical application, an option considered in practice is to divide a sample of size 1.5 million into 9 clusters.

5.2 Empirical simulations

We follow Shephard (2020) and provide simulation results calibrated to arithmetic returns from the SPDR S&P 500 ETF Trust (SPY), using data from August 1st 2018 to August 4th 2020. We simulate data from the following DGP:

$$\begin{aligned} y_i &= (z_i - \psi)\beta_1 + \varepsilon_i \\ z_i &= \psi + V_\nu \sigma_Z \sqrt{\frac{\nu-2}{\nu}},\ V_\nu \sim t_\nu \\ \varepsilon_i &\sim \mathcal{N}(0, (1 + |z_i - \psi|^\gamma) C^2) \end{aligned} \tag{58}$$

Here, z_i are calibrated to match the weekly returns. We set $\psi = 0.21$ to match the average weekly returns, and likewise set $\sigma_Z = 3.24$ to match the sample standard deviation. Fitting the normalized z_i to a student-t distribution leads to an implied estimate of ν of 2.16; as in Shephard (2020), we set this to 2.4, for slightly different reasons. Unlike that paper, we include conditional heteroskedasticity. The form of (58) is motivated so that for

some choices of γ, the standard Wald statistic based on the OLS estimator and the Huber-Eicker-White variance estimator leads to asymptotically valid inference under the null. We choose $\gamma = 0.3$ with this in mind. C is chosen so that $\mathbb{E}\,\varepsilon_i^2 = 4$, as in Shephard (2020). We set $\beta_{1,0} = 1$. To remove issues with the Wald-statistic based on the OLS estimator having poor size control, we first compute the finite-sample critical values to lead to valid inference. We set the sample size to $n = 100$.

We use two different test statistics to test the point null $H_0 : \beta_1 = 1$. The first standard heteroskedastic-robust Wald test using the OLS estimator, denoting the test statistic W_{OLS}. The second method is the IV method proposed by Shephard (2020). The estimator used is:

$$\hat{\beta}_{IV} := \frac{\sum_{i=1}^n \operatorname{sgn}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n |z_i - \bar{z}|}$$

We then construct the Wald test statistic W_{IV}. This estimator corresponds to a choice of $\psi = \operatorname{sgn}$ in Section 3.2.

Figure 2 plots the estimated power for a two sided test that $\beta_1 = 1.0$ based on 10000 Monte Carlo draws. On the left we see power, and on the right we see the power loss from using W_{OLS} instead of W_{IV}, on both an absolute and relative scale.

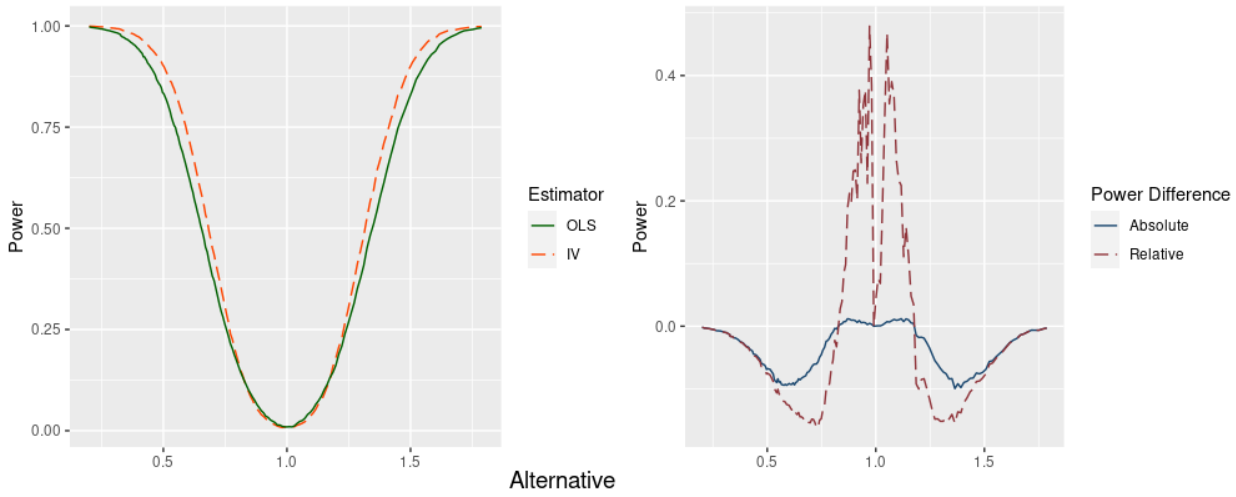

Figure 2: Comparison of OLS and IV estimators for heavy-tailed data

We highlight here that there is potentially a tradeoff between region in the alternative space where high-power is desired. In this context, the asymptotic variance of $\hat{\beta}_{OLS}$ is 1.2759, and thee asymptotic variance of $\hat{\beta}_{IV}$ is 1.4973. Local to the null $\beta_{1,0} = 1$, we see that the traditional test-statistic outperforms the Wald statistic based on $\hat{\beta}_{IV}$. This is consistent with traditional comparisons based on local asymptotics, but notice that the absolute different is fairly small. Farther from the null, the test statistic based on $\hat{\beta}_{IV}$ does better, which is

where we expect our theory to provide better predictions.

6 Application: Clustering, County, State, Region

As an application, we use American Community Survey (ACS) data from 2005-2019 to estimate the effect of minimum wage on employment. These data were used in MacKinnon et al. (2020a) to demonstrate how to apply their sequential testing procedure when trying to determine the appropriate clustering level. In this individual-level data set, natural clusters include state-year, state, year, and U.S. Census Division, hence referred to as "region." Our specification of interest is:

$$y_{ist} = \mu + \theta \text{mw}_{st} + z_{ist}'\gamma + \delta_t \text{year}_t + \delta_s \text{year}_s + \varepsilon_{ist} \tag{59}$$

for individual i in state s in year t. Here, y_{ist} is a binary variable equal to 1 if an individual is employed, 0 if unemployed. The parameter we will focus on is θ, the coefficient on the minimum wage (mw_{st}) in state s and year t. Other controls include individual level controls z_{ist}, which includes race, gender, age, and education dummies. We also include state and year fixed effects. The minimum wage data come from Neumark (2019). Since the ACS is collected annually, we collapse the minimum wage data to state-year averages. We also follow MacKinnon et al. (2020a) and restrict our attention to teenagers aged 16-19, keeping only individuals who are "children" of the respondent that have never been married. Individuals who have completed a year of college by the time they are 16 are dropped, as are those that report working in excess of 60 hours a week, on average. We focus our attention on individuals identifying as black or white.

Our approach is to take the approximation in (56) and use the approximating normal distribution to obtain power curves, as a function of Δ. We note that we must estimate β, V_n^A, ν_n, and Ξ_n^A. $\hat{\beta}$ is fixed across all test statistics. Since we assume that the finer cluster level is the correct cluster in all cases, we use the variance estimator at the finer level to estimate V_n^A and ν_n. The challenge becomes estimating Ξ_n^A. We must assume a certain kind of homogeneity across clusters. Let $X_{g(d)}$, $\varepsilon_{g(d)}$ be the finest cluster-level design matrix and error vector. When we assume this is the correct clustering level, we also assume that the cluster-sums $\sum_g X_{g(d)}'\varepsilon_{g(d)}$, $\sum_g X_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'X_{g(d)}$, $\sum_g X_{g(d)}'X_{g(d)}$ are i.i.d. across clusters. This will lead to a conservative estimate of the differences in power, as our asymptotic comparison implies that increased heterogeneity in the sizes of the too-coarse clusters leads to a large penalty for too-coarse clustering.

The reason we need this homogeneity is that we end up needing to estimate a variance

of the variance estimator, of the form:

$$\sum_{g=1}^{G} \mathbb{E}(X_g'\varepsilon)^4 - \left(\mathbb{E}(X_g'\varepsilon_g)^2\right)^2 \tag{60}$$

Plugging in the residuals for ε_g here, without assuming any kinds of similarity across g leads to this term being numerically 0.

We consider two base levels of correct clustering: state-year and state. The OLS estimate of θ is -0.00367, corresponding to the dashed blue line on each plot. We provide this as a guideline for where the empirically relevant portion of the power curves are. In Figure 3, we treat state-year as the true cluster level. In the left panel, we show the absolute power curves, with state-year and state seeming very close together, with region a bit below. On the right, we plot the difference between power curves, subtracting from the power of the test that uses state-year as the cluster variable. We see that that the predicted power loss of clustering at the region level is between 0.10 and 0.14 in the vicinity of the estimate of θ. Clustering at the state level does not induce nearly so large a penalty, with more modest power losses of 0.02-0.03. In Figure 4, states are the true clusters, but observations are independent across states within a region. We see a power loss of around 0.075 in the vicinity of $\hat{\theta}$.

These results and our methods complement those in MacKinnon et al. (2020a), where they found that clustering in this example should most likely be done at the state level. Our analysis focuses on the degree to which (block)-diagonal elements of the error-covariance matrix are sufficiently large to make coarse clustering too conservative, while they look to the off-(block) diagonal entries to ascertain validity. Together, the evidence points to clustering at the state level, especially since power losses in the case that state-year combinations are independent within state should be modest.

7 Conclusion

In this paper we develop a first-order asymptotic theory of Wald test statistics under fixed alternatives. We motivate this discussion by mapping the asymptotic distribution to a relative efficiency measure. Our main finding is that this alternative asymptotic framework distinguishes between approaches to testing that more classical approaches cannot order. This opens up the possibility of comparing different variance estimators that have previously been chosen based on simulation evidence, higher-order comparisons, or finite-sample criteria. Our approach applies to a broad class of models. One conclusion of particular interest for applied researchers is that there is an asymptotic cost for clustering at too-coarse a level. Our analysis also provides new insights into problems in econometric inference. Two notable

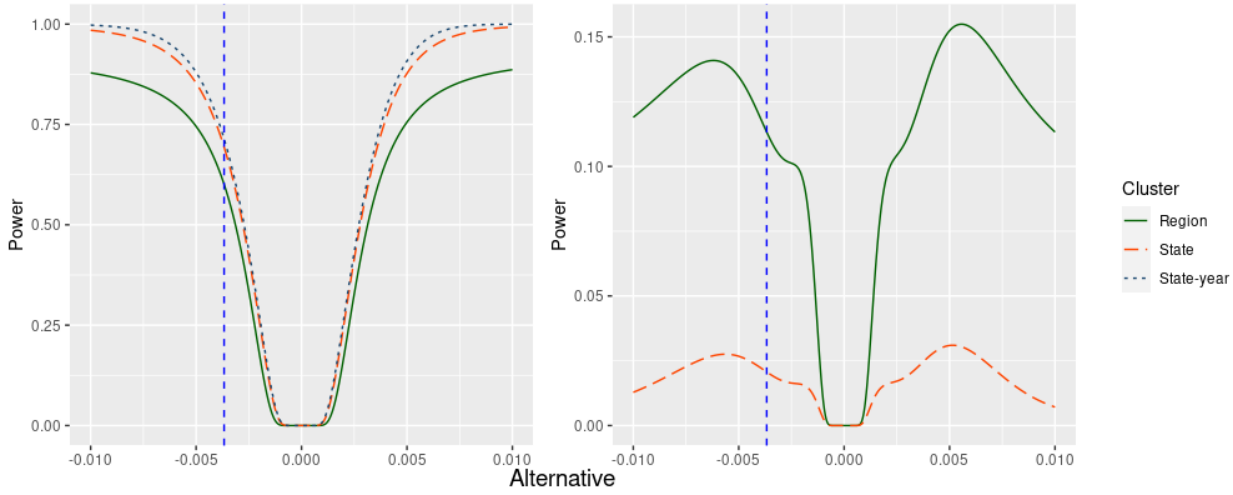

Figure 3: True clustering: state-year

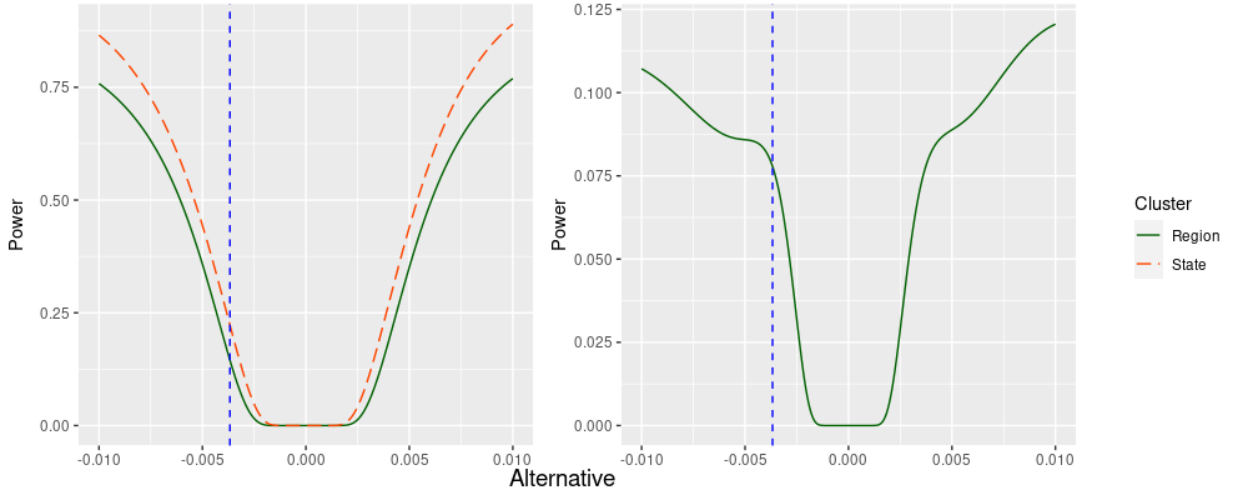

Figure 4: True clustering: state

examples are the consequences for power of heavy-tailed regressors/instruments, and issues arising from heavy-tailed errors in the first-stage regression in IV models.

There are also plenty of cases of interest not considered here. Our analysis could be applied to comparing commonly used heteroskedastic-robust variance estimators. We also did not pursue any high-dimensional or machine learning applications here, and it would be interesting to consider how our efficiency analysis could provide guidance for tuning parameter choices in that setting.

A Proofs of Main Results

Throughout, C will denote an arbitrary constant satisfying an upper bound, and λ will denote an arbitrary constant satisfying a lower bound; these will change based on the context.

A.1 Proof of Theorem 1

Under the assumptions of the theorem, for any matrix $A \in \mathbb{R}^{q\times q}$, we have that the estimator $\hat{\beta}$ is asymptotically linear:

$$\sqrt{n}(\hat{\beta}-\beta) = VQ'\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\beta) + o_P(1) \tag{61}$$

We then obtain an asymptotically linear form for $\hat{\Omega}$:

$$\begin{aligned}
\sqrt{n}\,\mathrm{tr}(A(\hat{\Omega}-\Omega)) &= \frac{1}{\sqrt{n}}\sum_i g(X_i,\hat{\beta})'Ag(X_i,\hat{\beta}) - \sqrt{n}\bar{g}(\hat{\beta})'A\bar{g}(\hat{\beta}) - \sqrt{n}\,\mathrm{tr}(A\Omega) && (62)\\
&= \frac{1}{\sqrt{n}}\sum_i g(X_i,\beta)'Ag(X_i,\beta) - \sqrt{n}\bar{g}(\beta)'A\bar{g}(\beta) - \sqrt{n}\,\mathrm{tr}(A\Omega) && (63)\\
&\quad + \mathbb{E}[g(X_i,\beta)'A\frac{\partial}{\partial\beta'}g(X_i,\beta)]\sqrt{n}(\hat{\beta}-\beta) + o_P(1) && (64)\\
&= \frac{1}{\sqrt{n}}\sum_i g(X_i,\beta)'Ag(X_i,\beta) - \mathbb{E}[g(X_i,\beta)'Ag(X_i,\beta)] && (65)\\
&\quad + \mathbb{E}[g(X_i,\beta)'A\frac{\partial}{\partial\beta'}g(X_i,\beta)]VQ'\Omega^{-1}\frac{1}{\sqrt{n}}\sum_i g(X_i,\beta) + o_P(1) && (66)
\end{aligned}$$

We also obtain an asymptotically linear form for $\hat{Q}$:

$$\sqrt{n}\,\mathrm{tr}(B(\hat{Q}-Q)) = \frac{1}{\sqrt{n}}\sum_i \mathrm{tr}(B\frac{\partial}{\partial\beta'}g(X_i,\hat{\beta})) - \sqrt{n}\,\mathrm{tr}(B'Q) \tag{67}$$

$$= \frac{1}{\sqrt{n}}\sum_i \sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\hat{\beta})'b_k - \sqrt{n}\,\mathrm{tr}(B'Q) \tag{68}$$

$$= \frac{1}{\sqrt{n}}\sum_i \sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\beta)'b_k \tag{69}$$

$$+ \left(\sum_{k=1}^{q} \mathbb{E}\left[\frac{\partial}{\partial\beta\partial\beta'}g_k(X_i,\beta)\right]b_k\right)' \sqrt{n}(\hat{\beta}-\beta) \tag{70}$$

$$- \sqrt{n}\,\mathrm{tr}(B'Q) \tag{71}$$

$$= \frac{1}{\sqrt{n}}\sum_i \sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\beta)'b_k - \mathbb{E}\left[\sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\beta)'b_k\right] \tag{72}$$

$$+ \left(\sum_{k=1}^{q} \mathbb{E}\left[\frac{\partial}{\partial\beta\partial\beta'}g_k(X_i,\beta)\right]b_k\right)' \sqrt{n}(\hat{\beta}-\beta) + o_P(1) \tag{73}$$

The relevant constants A, B, and c for us are:

$$A = \Omega^{-1}QV\ell\ell'VQ'\Omega^{-1} \tag{74}$$

$$= aa',\ a = \Omega^{-1}QV\ell \tag{75}$$

$$B = V\ell\ell'VQ'\Omega^{-1} \tag{76}$$

$$= ba',\ b = V\ell \tag{77}$$

$$c = \begin{pmatrix} \Omega^{-1}Q'V\left(\frac{-2\Delta}{\ell'V\ell}\ell - \frac{\Delta^2}{(\ell'V\ell)^2}\mathbb{E}[\frac{\partial}{\partial\beta'}g(X_i,\beta)Ag(X_i,\beta)] + \frac{2\Delta^2}{(\ell'V\ell)^2}\sum_{k=1}^{q}\mathbb{E}\left[\frac{\partial}{\partial\beta\partial\beta'}g_k(X_i,\beta)\right]b_k\right) \\ \frac{-\Delta^2}{(\ell'V\ell)^2} \\ \frac{2\Delta^2}{(\ell'V\ell)^2} \end{pmatrix} \tag{78}$$

and finally, Ξ is the asymptotic covariance matrix for:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} g(X_i,\beta) \\ (a'g(X_i,\beta))^2 - a'\Omega a \\ a'\frac{\partial}{\partial\beta'}g(X_i,\beta)b - a'Qb \end{pmatrix} \Rightarrow \mathcal{N}(0,\Xi) \tag{79}$$

We now list each partition Ξ_{ij}:

$$
\begin{aligned}
\Xi_{11} &= \Omega \\
\Xi_{12} = \Xi_{21}' &= \mathbb{E}[g(X_i,\beta)(a'g(X_i,\beta))^2] \\
\Xi_{22} &= \mathbb{E}(a'g(X_i,\beta))^4 - (a'\Omega a)^2 \\
\Xi_{23} = \Xi_{32} &= \mathbb{E}[(a'g(X_i,\beta))^2 a'\partial_{\beta'} g(X_i,\beta)b] - a'Qb\ell'b \\
\Xi_{33} &= \mathbb{E}(a'\partial_{\beta'} g(X_i,\beta)b)^2 - (a'Qb)^2
\end{aligned}
$$

The result then follows from Assumptions 3.1-3.3 and (22)-(25), and applying the standard multivariate central limit theorem.

A.2 Proof of Theorem 2

The test statistic can be expanded as:

$$
\frac{(\hat{\theta}-\theta_0)^2\hat{Q}^2}{\hat{\Omega}} - \frac{\Delta^2 Q^2}{\Omega} = \frac{(\hat{\theta}-\theta)^2\hat{Q}^2}{\hat{\Omega}} \tag{80}
$$

$$
+ \frac{2\Delta\hat{Q}^2(\hat{\theta}-\theta)}{\hat{\Omega}} \tag{81}
$$

$$
+ \frac{\Delta^2}{\hat{\Omega}}(\hat{Q}^2 - Q^2) \tag{82}
$$

$$
+ \Delta^2 Q^2\left(\frac{1}{\hat{\Omega}} - \frac{1}{\Omega}\right) \tag{83}
$$

(83), by the mean value theorem, can be written as:

$$
-\frac{\Delta^2 Q^2}{\tilde{\Omega}^2}(\hat{\Omega}-\Omega) \tag{84}
$$

$$
= -\frac{\Delta^2 Q^2}{\tilde{\Omega}^2}\left(\frac{1}{n}\sum_{i=1}^n z_i^2\varepsilon_i^2 - \Omega - \frac{2}{n}\sum_{i=1}^n z_i^2 x_i\varepsilon_i(\hat{\theta}-\theta) + \frac{1}{n}\sum_{i=1}^n z_i^2 x_i^2(\hat{\theta}-\theta)^2\right) \tag{85}
$$

The first term inside the parentheses is $O_P(n^{1-1/\gamma})$, by Assumption 3.5. Thus, consider, using Assumption 3.4 and the Marcinkiewicz-Zygmund law of large numbers (see e.g.

Korchevsky (2015)) FIX: center $z_i x_i$ by $\mathbb{E}\, z_i x_i$ first.

$$\left\| \frac{2}{n^{1/\gamma}} \sum_{i=1}^{n} z_i^2 x_i \varepsilon_i (\hat{\theta} - \theta) \right\| \leq \|\sqrt{n}(\hat{\theta} - \theta)\| \| n^{-1/2} \sum_{i=1}^{n} z_i \varepsilon_i \| \| n^{-1/\gamma} \sum_{i=1}^{n} (z_i x_i - \mathbb{E}\, z_i x_i) \| \| \mathbb{E}\, z_i x_i \| \tag{86}$$

$$= O_P(1) O_P(1) o_P(1) \tag{87}$$

Similarly, for the last term:

$$n^{-1/\gamma} \sum_{i=1}^{n} z_i^2 x_i^2 (\hat{\theta} - \theta)^2 = n^{-1-1/\gamma} \sum_{i=1}^{n} z_i^2 x_i^2 (\sqrt{n}(\hat{\theta} - \theta))^2 = o_P(1) O_P(1) \tag{88}$$

since for $\gamma \in (1, 2)$, $(1 + 1/\gamma)^{-1} < \gamma/2$. In the same manner, (82) is $o_P(n^{1-1/\gamma})$, and in fact (81) is $O_P(1/\sqrt{n})$. Then the result follows by the Generalized Central Limit Theorem, see e.g. Nolan (2020).

A.3 Proof of Theorem 3

For each test statistic, define:

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \hat{\theta})^2 \tag{89}$$

$$\tilde{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} ((y_i - z_i \hat{\pi} \hat{\theta}) - (x_i - z_i \hat{\pi}) \theta_0)^2 \tag{90}$$

Expand Wald first:

$$W_n - \frac{\Delta^2 Q^2}{\sigma^2} = \frac{(\hat{\theta} - \theta)^2 \hat{Q}^2}{\hat{\sigma}^2} \tag{91}$$

$$+ \frac{2\Delta \hat{Q}^2 (\hat{\theta} - \theta)}{\hat{\sigma}^2} \tag{92}$$

$$+ \frac{2\Delta^2 \bar{Q}}{\hat{\sigma}^2} (\hat{Q} - Q) \tag{93}$$

$$- \frac{\Delta^2 \hat{Q}^2}{\bar{\sigma}^4} (\hat{\sigma}^2 - \sigma^2) \tag{94}$$

Under the assumption of homoskedasticity, and $\mathbb{E}\, \varepsilon_i^4 < \infty$, $\mathbb{E}\, z_i^2 < \infty$, $\mathbb{E}\, \upsilon_i^2 < \infty$, (92)-(94)

are asymptotically normal at the usual $\sqrt{n}$ rate. Now, consider the AR test:

$$R_n - \frac{\Delta^2 Q^2}{\sigma^2 + 2\Delta\rho + \Delta^2\sigma_v^2} = \frac{(\hat{\theta} - \theta)^2 \hat{Q}^2}{\tilde{\sigma}^2} \tag{95}$$

$$+ \frac{2\Delta \hat{Q}^2 (\hat{\theta} - \theta)}{\tilde{\sigma}^2} \tag{96}$$

$$+ \frac{2\Delta^2 \bar{Q}}{\tilde{\sigma}^2}(\hat{Q} - Q) \tag{97}$$

$$- \frac{\Delta^2 \hat{Q}^2}{\bar{\sigma}^2 + 2\Delta\bar{\rho} + \Delta^2\bar{\sigma}_v^2}(\tilde{\sigma}^2 - \sigma^2 - 2\Delta\rho - \Delta^2\sigma_v^2) \tag{98}$$

The terms (96) and (97) are still $\sqrt{n}$-normal. Now, the last term is dominated by the terms involving v_i^2:

$$\tilde{\sigma}^2 - \sigma^2 - 2\Delta\rho - \Delta^2\sigma_v^2 = \frac{1}{n}\sum_{i=1}^{n} (\varepsilon_i + \Delta v_i + z_i(\hat{\pi} - \pi)\Delta - z_i\hat{\pi}(\hat{\theta} - \theta))^2 - \sigma^2 - 2\Delta\rho - \Delta^2\sigma_v^2 \tag{99}$$

$$= \Delta^2 \frac{1}{n}\sum_{i=1}^{n} v_i^2 - \sigma_v^2 \tag{100}$$

$$+ 2\Delta \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i v_i - \rho \tag{101}$$

$$+ O_P(1/\sqrt{n}) \tag{102}$$

with the remaining terms of a similar order to those handled in the case of the Wald statistic. Now, $\mathbb{E}\,|\varepsilon_i v_i|^\gamma < \infty$, since $\mathbb{E}\,\varepsilon_i^4 < \infty$ and $\mathbb{E}\,v_i^2 < \infty$, so we can use Hölder's inequality and the Marcinkiewicz-Zygmund law of large numbers, so (101) is negligible when normalized by $n^{1-1/\gamma}$. Thus, we are left with (100), which by assumption converges to a stable law, and the result follows.

A.4 Proof of Theorem 4

We proceed in a similar fashion to Kato (2012). Expanding the test-statistic:

$$W_n - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega Q_\tau^{-1} \ell} \tag{103}$$

$$= \frac{(\ell' \hat{\beta}(\tau) - \theta)^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega} \hat{Q}_\tau^{-1} \ell} \tag{104}$$

$$+ \frac{2\Delta(\ell' \hat{\beta}(\tau) - \theta)}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega} \hat{Q}_\tau^{-1} \ell} \tag{105}$$

$$+ \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega} \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega Q_\tau^{-1} \ell} \tag{106}$$

By Slutksy's theorem, Assumption 3.7 and the bandwidth condition in Assumption 3.10, (104) is $O_P(1/n)$ and (105) is $O_P(1/\sqrt{n})$. Turning to (106), consider the expansion:

$$\frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell} \tag{107}$$

$$= \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \Omega_\tau \hat{Q}_\tau^{-1} \ell} + \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell} \tag{108}$$

We note that the first difference is of the same order as $\|\hat{\Omega}_\tau - \Omega_\tau\|$, and therefore is of order $O_P(1/\sqrt{n})$. We turn to the second difference, and note by the mean value theorem:

$$\frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell} = \operatorname{tr}(A(\hat{Q}_\tau - Q_\tau)) + o_P(1/\sqrt{n h_n}) \tag{109}$$

$$A := \frac{2Q^{-1}\Delta^2 \ell \ell' Q^{-1} \Omega Q^{-1}}{(\ell' Q^{-1} \Omega Q^{-1} \ell)^2} \tag{110}$$

We will supply an argument for the rate assertion later. Let $\psi_\beta(z_i) := \operatorname{tr}(A x_i x_i') K((y - x_i'\beta)/h_n)$. Thus, we start by considering that:

$$h_n^{-1} \mathbb{P}_n \psi_{\hat{\beta}}(z_i) \xrightarrow{P} \operatorname{tr}(A \, \mathbb{E}[x_i x_i' f(0|x_i)])$$

but that since $h_n^{-1} \mathbb{E} \psi_\beta(z_i) \neq \operatorname{tr}(A \, \mathbb{E}[x_i x_i' f(0|x_i)])$, and generally the difference is asymptotically non-negligible, we start by looking at the simple decomposition:

$$\sqrt{\frac{n}{h_n}}\left(\mathbb{P}_n\,\psi_{\hat{\beta}} - \mathbb{E}\,\psi_\beta\right) = h_n^{-1/2}\left(\mathbb{G}_n\,\psi_b|_{b=\hat{\beta}} - \mathbb{G}_n\,\psi_\beta\right) \tag{111}$$

$$+\, h_n^{-1/2}\,\mathbb{G}_n\,\psi_\beta \tag{112}$$

$$+\, \sqrt{nh_n}\left(h_n^{-1}\,\mathbb{E}\,\psi_b(z_i)|_{b=\hat{\beta}} - h_n^{-1}\,\mathbb{E}\,\psi_\beta(z_i)\right) \tag{113}$$

(113) can be bounded since:

$$\mathbb{E}\,h_n^{-1}\psi_b(z_i) = \mathbb{E}\left[x_i'Ax_i(f(x_i'(\beta_\tau - b)|x_i) + \frac{1}{2}f''(x_i'(\beta_\tau - b)|x_i)h_n^2 + o(h_n^2))\right] \tag{114}$$

Thus, by Assumption 3.7 and 3.9, (113) is $o_P(1)$. Thus, we can turn our attention to (111), since (112) will satisfy a standard Lindeberg CLT for kernel density estimators. An implication of Assumption 3.8 is that there exist functions K_1, K_2 such that K_i is non-negative, non-decreasing, and $K = K_1 - K_2$. Furthermore, $|K|_v = |K_1|_v + |K_2|_v$, so we have a simple form of the total-variation norm. Using arguments similar to those found in Einmahl and Mason (2000), we have that, for t, $s \in \mathbb{R}^p$, letting $\delta_t = t - \beta$, $\delta_s = s - \beta$,

$$\begin{aligned} K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right) &= K_1\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K_1\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right) \\ &\quad - \left(K_2\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K_2\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right) \\ &= \int_{\frac{\varepsilon_i - x_i'\delta_s}{h_n}}^{\frac{\varepsilon_i - x_i'\delta_t}{h_n}} dK_1(x) - \int_{\frac{\varepsilon_i - x_i'\delta_s}{h_n}}^{\frac{\varepsilon_i - x_i'\delta_t}{h_n}} dK_2(x) \end{aligned}$$

This implies, via the triangle inequality,

$$\left|K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right| \leq \int \left|\mathbb{1}_{[\frac{\varepsilon_i - x_i'\delta_s}{h_n},\, \frac{\varepsilon_i - x_i'\delta_t}{h_n}]}(x)\right| d(K_1(x) + K_2(x)) \tag{115}$$

Thus, using (115), we can use Hölder's inequality to bound the mean-squared difference:

$$\mathbb{E}\left[\left(K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right)^2 \middle| x_i\right] \leq \int \mathbb{E}\left|\mathbb{1}_{[\frac{\varepsilon_i - x_i'\delta_s}{h_n}, \frac{\varepsilon_i - x_i'\delta_t}{h_n}]}(x)\right| d(K_1(x) + K_2(x))|K|_v \tag{116}$$

$$= \int \left|\int_{x_i'\delta_s + h_n x}^{x_i'\delta_t + h_n x} f(\varepsilon|x_i) d\varepsilon\right| d(K_1(x) + K_2(x))|K|_v \tag{117}$$

$$\leq \|f(\cdot|x_i)\|_\infty |K|_v^2 \|x_i\|_2 \|t - s\|_2 \tag{118}$$

Now, by Assumption 3.9:

$$\mathbb{E}\left[\left(K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right)^2\right] = O(\|t - s\|_2) \tag{119}$$

Now, we return to (111). For any $\delta > 0$, let $N_{\delta/\sqrt{n}}(\beta)\}$ be a $\delta/\sqrt{n}$ neighborhood of β. Then, we have that for any $\epsilon > 0$,

$$P\left(\sup_{b \in N_{\delta/\sqrt{n}}(\beta)} |\mathbb{G}_n(\psi_b - \psi_\beta)| > h_n^{1/2}\epsilon\right) \leq \frac{1}{\epsilon h_n^{1/2}} \mathbb{E}\left(\sup_{b \in N_{\delta/\sqrt{n}}(\beta)} |\,\mathbb{G}_n(\psi_b - \psi_\beta)|\right)$$

We now need a slight extension of a VC-class result from Giné and Nickl (2016):

Proposition 2. *Let $\mathcal{K} = \{(\varepsilon,\, x) \mapsto K\left(\frac{\varepsilon - x't}{h}\right) : t \in \mathbb{R}^p,\, h > 0\}$. Then $\mathcal{K}$ is of VC-subgraph type.*

We are now ready to use the maximal inequality of Chernozhukov et al. (2014):

$$h_n^{-1/2}\, \mathbb{E}\left(\sup_{b \in N_{\delta/\sqrt{n}}(\beta)} |\,\mathbb{G}_n(\psi_b - \psi_\beta)|\right) = O\left(\sqrt{\frac{\log n}{h_n n^{1/2}}}\right)$$

where in the notation of Corollary 5.1 of Chernozhukov et al. (2014), we can choose $\sigma^2 = O(1/\sqrt{n})$ by (119). This means that when $h_n = o(\log n/\sqrt{n})$, (111) converges to zero in probability.

Thus, we have that, by standard results on kernel density estimation,

$$\sqrt{nh_n}\left(W_n - \frac{\Delta^2}{\ell' Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} - \frac{1}{2}\,\mathbb{E}[x_i' A x_i f''(0|x_i)h_n^2]\right) \Rightarrow \mathcal{N}\left(0,\, \mathbb{E}[(x_i' A x_i)^2 f(0|x_i) R_K]\right) \tag{120}$$

A.5 Proof of Theorem 5

The first goal will be to derive a central limit theorem for a particular linear functionals of:

$$S_n := \frac{1}{n}\sum_{g=1}^{G}\begin{pmatrix} Z_g'\varepsilon_g \\ \ell' Q_n^{-1}(Z_g'\varepsilon_g\varepsilon_g' Z_g - \mathbb{E}[Z_g'\varepsilon_g\varepsilon_g' Z_g])(Q_n')^{-1}\ell \\ \ell' Q_n^{-1}(Z_g' X_g - \mathbb{E}\, Z_g' X_g) Q_n^{-1}\Omega_n (Q_n')^{-1}\ell \end{pmatrix} \tag{121}$$

Define:

$$c_n := \frac{1}{n}\sum_{g=1}^{G}\mathbb{E}[X_g' Z_g a_n a_n' Z_g'\varepsilon_g] \tag{122}$$

$$\xi_n := \Delta/\ell' V_n \ell \tag{123}$$

$$Y_g := \frac{1}{n}\begin{pmatrix} Z_g'\varepsilon_g \\ a_n'(Z_g'\varepsilon_g\varepsilon_g' Z_g - \mathbb{E}[Z_g'\varepsilon_g\varepsilon_g' Z_g])a_n \\ a_n'(Z_g' X_g - \mathbb{E}\, Z_g' X_g) b_n \end{pmatrix} \tag{124}$$

The linear combination we will be interested in is:

$$\nu_n := \begin{pmatrix} 2(\xi_n a_n - (Q_n')^{-1} c_n) \\ -\xi_n^2 \\ 2\xi_n^2 \end{pmatrix} \tag{125}$$

First, note that by Assumption 4.2, $\Xi_n^G := \mathbb{E}\, S_n S_n'$ exists and is well-behaved, and therefore if we are going to find a limit distribution, we would expect it to be:

$$(\nu_n' \Xi_n^G \nu_n)^{-1/2}\nu_n' S_n \Rightarrow \mathcal{N}(0,1) \tag{126}$$

By Assumption 4.3, the Y_g are uniformly integrable, and thus Assumptions 4.1-4.2, the assumptions of Corollary 1 in Hansen and Lee (2019) are satisfied and (126) holds. Furthermore, since $S_n \xrightarrow{P} 0$, Ξ_n^G contains the information about the rate of convergence of the elements of S_n. Next, note that:

$$W_n^G - \frac{\Delta^2}{\ell' V_n \ell} \xrightarrow{P} 0 \tag{127}$$

To see why this is, examine the three components of the difference:

$$W_n^G - \frac{\Delta^2}{\ell' V_n \ell} = \frac{(\ell'\hat{\beta} - \theta_0)^2}{\ell' \hat{V}_n^G \ell} - \frac{\Delta^2}{\ell' V_n \ell} \tag{128}$$

$$= \frac{(\ell'\hat{\beta} - \theta)^2}{\ell' \hat{V}_n^G \ell} \tag{129}$$

$$+ \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell' \hat{V}_n^G \ell} \tag{130}$$

$$+ \Delta^2 \left(\frac{1}{\ell' \hat{V}_n^G \ell} - \frac{1}{\ell' V_n \ell} \right) \tag{131}$$

The first term is $O_P(1/n)$, by Theorem 9 in Hansen and Lee (2019). The third term is $o_P(1)$ by the continuous mapping theorem, the rank condition, and Theorem 9 in Hansen and Lee (2019). Lastly, the middle term is equal to:

$$\frac{2\Delta}{\sqrt{n}\sqrt{\ell' V_n \ell}} \frac{\sqrt{n}(\ell'\hat{\beta} - \theta)}{\sqrt{\ell' V_n \ell}} + o_P(1) \tag{132}$$

This term is $O_P(1)$:

$$\frac{\sqrt{n}(\ell'\hat{\beta} - \theta)}{\sqrt{\ell' V_n \ell}} \tag{133}$$

by Theorem 9 in Hansen and Lee (2019). Examining the other term, we have:

$$n\ell' V_n \ell = \ell' Q_n^{-1} \sum_{g=1}^{G} \mathbb{E}\, Z_g' \varepsilon_g \varepsilon_g' Z_g (Q_n')^{-1} \ell \tag{134}$$

The rank condition on Q_n implies that this term goes to ∞, therefore (132) converges to 0 in probability, and therefore $W_n^G - \Delta^2/\ell' V_n \ell \xrightarrow{P} 0$.

There are three rates of convergence at play here:

$$\ell'\hat{\beta} - \theta \xrightarrow{P} 0 \tag{135}$$

$$\|\hat{Q}_n - Q_n\| \xrightarrow{P} 0 \tag{136}$$

$$\|\hat{\Omega}_n - \Omega_n\| \xrightarrow{P} 0 \tag{137}$$

The rate of convergence of $\ell'\hat{\beta} - \theta$ will be the fastest rate, at least weakly. To see why, consider (132).

These terms play a role in how closely $\nu_n' S_n$ approximates the centered test statistic.

Rewriting:

$$W_n^G - \frac{\Delta^2}{\ell' V_n \ell} = \frac{(\ell'\hat{\beta} - \theta_0)^2}{\ell' \hat{V}_n^G \ell} - \frac{\Delta^2}{\ell' V_n \ell} \tag{138}$$

$$= \frac{(\ell'\hat{\beta} - \theta)^2}{\ell' \hat{V}_n^G \ell} \tag{139}$$

$$+ \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell' \hat{V}_n^G \ell} \tag{140}$$

$$+ \Delta^2 \left(\frac{1}{\ell' \hat{V}_n^G \ell} - \frac{1}{\ell' V_n \ell} \right) \tag{141}$$

(140) can be properly normalized to be asymptotically normal, so the main component of interest is (141). Using a Taylor expansion, we write:

$$\Delta^2 \left(\frac{1}{\ell' \hat{V}_n^A \ell} - \frac{1}{\ell' V_n^G \ell} \right) = -\frac{\Delta^2}{(\ell' \tilde{V}_n^G \ell)^2} \left(\ell' \hat{V}_n^G \ell - \ell' V_n \ell \right) \tag{142}$$

where $\tilde{V_n^G}$ is a convex combination of $\hat{V}_n^G$ and V_n, since we are using the scalar version of Taylor's theorem. Now, we separate (142) into a component depending on $\hat{Q}_n$ and a component depending on $\hat{\Omega}_n$:

$$\ell' \hat{V}_n^G \ell - \ell' V_n \ell = \ell' \hat{Q}_n^{-1} \hat{\Omega}_n^G (\hat{Q}_n')^{-1} \ell - \ell' Q_n^{-1} \Omega_n (Q_n')^{-1} \ell \tag{143}$$

$$= \ell' \hat{Q}_n^{-1} \hat{\Omega}_n^G (\hat{Q}_n')^{-1} \ell - \ell' \hat{Q}_n^{-1} \Omega_n (\hat{Q}_n')^{-1} \ell \tag{144}$$

$$+ \ell' \hat{Q}_n^{-1} \Omega_n (\hat{Q}_n')^{-1} \ell - \ell' Q_n^{-1} \Omega_n (Q_n')^{-1} \ell \tag{145}$$

First, consider (144). This terms is almost ready to analyze, but we are using a feasible estimator of Ω_n rather than the infeasible estimator with known ε_{gi}. Thus, we have:

$$\check{\Omega}_n^G := \frac{1}{n} \sum_{g=1}^{G} Z_g' \varepsilon_g \varepsilon_g' Z_g \tag{146}$$

$$\hat{\Omega}_n^G = \check{\Omega}_n^G - \frac{1}{n} \sum_{g=1}^{G} Z_g' \varepsilon_g (\hat{\beta} - \beta)' X_g' Z_g - \frac{1}{n} \sum_{g=1}^{G} Z_g' X_g (\hat{\beta} - \beta) \varepsilon_g' Z_g \tag{147}$$

$$+ \frac{1}{n} \sum_{g=1}^{G} Z_g' X_g (\hat{\beta} - \beta)(\hat{\beta} - \beta)' X_g' Z_g \tag{148}$$

We will actually show that we only need to consider (147), and (148) will be negligible, since

it is of lower-order. With our moment conditions, we have that there exists C such that:

$$\left\| \frac{1}{n} \sum_{g=1}^{G} Z_g' X_g (\hat{\beta} - \beta)(\hat{\beta} - \beta)' X_g' Z_g \right\| \leq C \|\hat{\beta} - \beta\|^2 \tag{149}$$

In fact, the slowest rate of convergence of this term is going to be $\check{\Omega}_n^G$, since

$$\left\| \frac{1}{n} \sum_{g=1}^{G} Z_g' \varepsilon_g (\hat{\beta} - \beta)' X_g' Z_g \right\| \leq C \|\hat{\beta} - \beta\| \tag{150}$$

We proceed by next carefully performing a Taylor expansion of (145).

$$\ell' \hat{Q}_n^{-1} \Omega_n (\hat{Q}_n')^{-1} \ell - \ell' Q_n^{-1} \Omega_n (Q_n')^{-1} \ell = \operatorname{tr}(-2(\tilde{Q}_n)^{-1} \Omega_n (\tilde{Q}_n')^{-1} \ell \ell' (\tilde{Q}_n)^{-1} (\hat{Q}_n - Q_n) \tag{151}$$

where $\tilde{Q}_n$ is an element-wise convex combination of Q_n and $\hat{Q}_n$, i.e. $[\tilde{Q}_n]_{ij} = \omega_{ij}[Q_n]_{ij} + (1 - \omega_{ij})[\hat{Q}_n]_{ij}$, for possibly different ω_{ij}. Gathering terms from (140), (144), and (151), and the constants in (142), we have:

$$\begin{aligned}
W_n^G - \frac{\Delta^2}{\ell' V_n^G \ell} &= \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell' \hat{V}_n \ell} && (152)\\
&\quad - \frac{\Delta^2}{(\ell' \tilde{V}_n^G \ell)^2} \operatorname{tr}((\hat{Q}_n')^{-1} \ell \ell' \hat{Q}_n^{-1} (\hat{\Omega}_n^G - \Omega_n)) && (153)\\
&\quad + \frac{2\Delta^2}{(\ell' \tilde{V}_n^G \ell)^2} \operatorname{tr}((\tilde{Q}_n)^{-1} \Omega_n (\tilde{Q}_n')^{-1} \ell \ell' (\tilde{Q}_n)^{-1} (\hat{Q}_n - Q_n) && (154)\\
&\quad + O_P(1/n) && (155)
\end{aligned}$$

This implies that the asymptotic distribution of $W_n^G - \frac{\Delta^2}{\ell' V_n^G \ell}$ should be the same as:

$$\begin{aligned}
\bar{W}_n^G - \frac{\Delta^2}{\ell' V_n \ell} &:= \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell' V_n \ell} - 2\frac{1}{n} \sum_{g=1}^{G} \mathbb{E}[\varepsilon_g' Z_g (Q_n')^{-1} \ell \ell' Q_n^{-1} Z_g' X_g](\hat{\beta} - \beta) && (156)\\
&\quad - \frac{\Delta^2}{(\ell' V_n \ell)^2} \operatorname{tr}((Q_n')^{-1} \ell \ell' Q_n^{-1} (\check{\Omega}_n^G - \Omega_n)) && (157)\\
&\quad + \frac{2\Delta^2}{(\ell' V_n \ell)^2} \operatorname{tr}((Q_n)^{-1} \Omega_n (Q_n')^{-1} \ell \ell' (Q_n)^{-1} (\hat{Q}_n - Q_n) && (158)
\end{aligned}$$

since

$$\|W_n^G - \bar{W}_n^G\| = o_P\left(\max\left\{\|\hat{\beta} - \beta\|, \|\hat{Q}_n - Q_n\|, \|\check{\Omega}_n - \Omega_n\|\right\}\right) \tag{159}$$

Furthermore, we also have that:

$$\|\bar{W}_n^G - \Delta^2/\ell' V_n \ell - \nu_n' S_n\| = o_P\left(\max\left\{\|\hat{\beta} - \beta\|, \|\hat{Q}_n - Q_n\|, \|\check{\Omega}_n - \Omega_n\|\right\}\right) \tag{160}$$

Thus, we have that:

$$(\nu_n' \Xi_n^G \nu_n)^{-1/2}\left(W_n^G - \frac{\Delta^2}{\ell' V_n \ell}\right) = (\nu_n' \Xi_n^G \nu_n)^{-1/2} \nu_n' S_n + o_P(1) \tag{161}$$

The proof when using $\hat{V}_n^D$ is similar. Now, when looking at $\Xi_n^D - \Xi_n^G$, we note that all terms are zero, except for the second-to-last diagonal element. We need to compute $[\Xi_n^D]_{q+1,q+1}$ in terms of the moments of $(Y_g)_{q+1,q+1}$

$$[\Xi_n^D]_{q+1,q+1} - [\Xi_n^G]_{q+1,q+1} = \frac{1}{n^2}\sum_{d=1}^{D}\sum_{g\neq h}\mathbb{E}(a_n' Z_{g(d)}' \varepsilon_{g(d)})^2\,\mathbb{E}(a_n' Z_{h(d)}' \varepsilon_{h(d)})^2 \geq 0 \tag{162}$$

To see why this is, consider the cumulants; we drop the subscripts for the element in Y_d and Y_g since it is clear what we mean here:

$$\begin{aligned}
Y_d &:= \sum_{g=1}^{G_d} a_n' Z_g' \varepsilon_g && (163)\\
&= \sum_{g=1}^{G_d} Y_{dg} && (164)\\
\mathbb{E}\, Y_d &= \sum_{g=1}^{G_d} \mathbb{E}\, Y_{g(d)} && (165)\\
&= 0 && (166)\\
\mathrm{Var}(Y_d) = \mathbb{E}\, Y_d^2 &= \sum_{g=1}^{G_d} \mathbb{E}(a_n' Z_g' \varepsilon)^2 && (167)\\
&= \sum_{g=1}^{G_d} \mathrm{Var}(Y_{dg}) && (168)
\end{aligned}$$

Let $k_4(X)$ be the 4th cumulant of X. Then, we have that:

$$\mathbb{E}\, Y_d^4 = k_4(Y_d) + 3(\mathbb{E}\, Y_d^2)^2 = k_4(Y_d) + 3\left(\sum_{g=1}^{G_d}\mathrm{Var}(Y_{g(d)})\right)^2 \tag{169}$$

By the properties of cumulants, we have that, using properties of the cumulants again:

$$\begin{aligned} k_4(Y_d) &= \sum_{g=1}^{G_d} k_4(Y_{g(d)}) \\ &= \sum_{g=1}^{G_d} \mathbb{E}\, Y_{g(d)}^4 - 3(\mathbb{E}\, Y_{g(d)}^2)^2 \\ \mathbb{E}\, Y_d^4 &= \sum_{g=1}^{G_d} \mathbb{E}\, Y_{g(d)}^4 + 3 \sum_{h \neq g} \operatorname{Var}(Y_{g(d)}) \operatorname{Var}(Y_{h(d)}) \end{aligned}$$

Thus:

$$[\Xi_n^D]_{q+1,q+1} - [\Xi_n^G]_{q+1,q+1} = \frac{1}{n^2} \sum_{d=1}^{D} \sum_{g \neq h} \mathbb{E}(Y_{g(d)})^2\, \mathbb{E}(Y_{h(d)})^2 \tag{170}$$

$$= \frac{1}{n^2} \sum_{d=1}^{D} \sum_{g \neq h} \mathbb{E}(a_n' Z_{g(d)}' \varepsilon_{g(d)})^2\, \mathbb{E}(a_n' Z_{h(d)}' \varepsilon_{h(d)})^2 \geq 0 \tag{171}$$

$$\tag{172}$$

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When Should You Adjust Standard Errors for Clustering?" Tech. rep., National Bureau of Economic Research.

Amemiya, T. (1982): "Two Stage Least Absolute Deviations Estimators," *Econometrica: Journal of the Econometric Society*, 689–711.

Anatolyev, S. (2019): "Many Instruments and/or Regressors: A Friendly Guide," *Journal of Economic Surveys*, 33, 689–726.

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Bahadur, R. R. (1967): "Rates of Convergence of Estimates and Test Statistics," *The Annals of Mathematical Statistics*, 38, 303–324.

Bentkus, V., B. Y. Jing, Q. M. Shao, and W. Zhou (2007): "Limiting Distributions of the Non-Central t-Statistic and Their Applications to the Power of t-Tests under Non-Normality," *Bernoulli*, 13, 346–364.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly journal of economics*, 119, 249–275.

Breiman, L. (1965): "On Some Limit Theorems Similar to the Arc-Sin Law," *Theory of Probability and its Applications*, 10, 323–331.

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. and T. Otsu (2012): "Hodges–Lehmann Optimality for Testing Moment Conditions," *Journal of Econometrics*, 171, 45–53.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Gaussian Approximation of Suprema of Empirical Processes," *The Annals of Statistics*, 42, 1564–1597.

Chow, T. L. and J. L. Teugels (1978): "The Sum and the Maximum of Iid Random Variables," in *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 21–25.

Dembo, A. and O. Zeitouni (2009): *Large Deviations Techniques and Applications*, Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg.

Durrett, R. (2019): *Probability: Theory and Examples*, vol. 49, Cambridge university press.

Einmahl, U. and D. M. Mason (2000): "An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators," *Journal of Theoretical Probability*, 13, 1–37.

Engle, R. F. (1984): "Chapter 13 Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics*, Elsevier, vol. 2, 775–826.

Giné, E. and R. Nickl (2016): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge University Press.

Hansen, B. E. and S. Lee (2019): "Asymptotic Theory for Clustered Samples," *Journal of econometrics*, 210, 268–290.

Hansen, C. B. (2007): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large," *Journal of Econometrics*, 141, 597–620.

Hettmansperger, T. P. (1972): "Hodges and Lehmann Approximate Efficiency," 279–286.

Hodges, J. and E. Lehmann (1956): "The Efficiency of Some Nonparametric Competitors of the T-Test," *Annals of Mathematical Statistics*, 27, 324–335.

Honoré, B. E. and L. Hu (2004): "On the Performance of Some Robust Instrumental Variables Estimators," *Journal of Business & Economic Statistics*, 22, 30–39.

Jiao, X. (2019): "A Simple Robust Procedure in Instrumental Variables Regression," .

Kato, K. (2012): "Asymptotic Normality of Powell's Kernel Estimator," *Annals of the Institute of Statistical Mathematics*, 64, 255–273.

Kim, D. and P. Perron (2009): "Assessing the Relative Power of Structural Break Tests Using a Framework Based on the Approximate Bahadur Slope," *Journal of Econometrics*, 149, 26–51.

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica: journal of the Econometric Society*, 33–50.

Korchevsky, V. (2015): "Marcinkiewicz-Zygmund Strong Law of Large Numbers for Pairwise i.i.d. Random Variables," .

MacKinnon, J. G., M. Nielsen, M. D. Webb, et al. (2020a): "Cluster-Robust Inference: A Guide to Empirical Practice," Tech. rep., Qed working paper, Queen's University.

MacKinnon, J. G., M. Ø. Nielsen, and M. Webb (2020b): "Testing for the Appropriate Level of Clustering in Linear Regression Models," Tech. rep., Queen's Economics Department Working Paper.

Müller, U. K. (2020): "A More Robust T-Test," *arXiv preprint arXiv:2007.07065*.

Neumark, D. (2019): "Minimum Wage Data," .

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of econometrics*, 4, 2111–2245.

Nolan, J. P. (2020): *Univariate Stable Distributions*, Springer.

Omey, E. and S. van Gulck (2009): "Domains of Attraction of the Real Random Vector (x,X2) and Applications," *Publications de l'Institut Mathematique*, 86, 41–53.

Pitman, E. J. (1949): "Notes on Non-Parametric Statistical Inference," Tech. rep., North Carolina State University. Dept. of Statistics.

Powell, J. L. (1983): "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," *Econometrica: Journal of the Econometric Society*, 1569–1575.

——— (1991): "Estimation of Monotonic Regression Models under Quantile Restrictions," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics : Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, Cambridge [England] ; New York : Cambridge University Press, 1991.

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

Sasaki, Y. and Y. Wang (2020): "Testing Finite Moment Conditions for the Consistency and the Root-n Asymptotic Normality of the GMM and m Estimators," *arXiv preprint arXiv:2006.02541*.

Shao, Q. and R. Zhang (2009): "Asymptotic Distributions of Non-Central Studentized Statistics," *Science in China, Series A: Mathematics*, 52, 1262–1284.

Shephard, N. (2020): "An Estimator for Predictive Regression : Reliable Inference for Financial Economics," .

Sølvsten, M. (2020): "Robust Estimation with Many Instruments," *Journal of Econometrics*, 214, 495–512.

Staiger, D. O. and J. H. Stock (1994): "Instrumental Variables Regression with Weak Instruments," .

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.

Young, A. (2021): "Leverage, Heteroskedasticity and Instrumental Variables in Practical Application," .