Comparing Variance Estimators: a Test-Based Relative-Efficiency Approach *

Samuel P. Engle†

Job Market Paper

University of Wisconsin-Madison
Department of Economics

This version: January 1, 2022
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 of hypothesis tests indicate all consistent variance estimators lead to equivalent Wald tests. This paper develops a simple relative efficiency measure which leads to several new conclusions. These include quantifying the power loss associated with using cluster-robust variance estimators when using overly coarse clusters, recommending particular kernels for estimating the asymptotic variance in quantile regression, and comparing the power of Anderson-Rubin tests to the standard Wald test. As a byproduct, the asymptotic distributions of several test statistics are derived under fixed alternatives. Simulation evidence indicates the new asymptotic efficiency measure provides good finite-sample predictions. In an application using data from the American Community Survey, it is demonstrated how to use the new approach for conducting power analysis when looking at the effect of minimum wage increases on employment.

*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, Eric Auerbach, Kei Hirano, Bo Honore, Jim Powell, Yuya Sasaki, Ken West, Annie Lee, John Stromme, Anna Trubnikova, Anson Zhou, 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. 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.

The choice of variance estimator is an every present decision in applied work. There is a menu of available robust consistent variance estimators in standard statistical packages. Researchers with grouped data must determine whether to compute cluster-robust standard errors, and what level to cluster at. In the case of quantile regression, researchers choose a kernel density estimator to use. In likelihood settings under correct specification, the Fisher information matrix can be estimated using the outer product of the score or the second derivative of the log-likelihood. Robust variance estimators in time series involve choosing a kernel and truncation point. We will not consider all these examples here, however we provide a framework that is suited to studying the effect of variance estimation in many of these contexts.

This paper makes three key contributions to the econometric literature on hypothesis testing. First, we develop a new approach for conducting power analysis in a wide class of econometric models. This approach leads to a new way to compare the relative efficiency of tests that is sensitive to the choice of variance estimator. Second, we apply this approach to several applications, leading to new insights into testing in these environments: cluster-robust inference, quantile regression, and linear instrumental variables (IV) models. Last, as an intermediate step of possibly independent interest, we derive the asymptotic distribution of Wald test statistics under fixed-alternatives in these different econometric settings.

Considering the behavior of test statistics under fixed alternatives is a key part of how we distinguish between different variance estimators. 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 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 parameter estimator is used in two different tests, even if different

consistent variance estimators are used. This equivalence no longer holds in our asymptotic theory when different estimators of the asymptotic variance are used. To compare these test statistics, we propose using an asymptotic relative efficiency (ARE) measure which compares tests under a regime where size converges to 0 and power converges to a constant in $(1/2, 1)$. Our approach contrasts with the local ARE of Pitman (1949), the most commonly used approach in econometrics. The ARE measure we propose can be compared to the measure proposed in Bahadur (1967). A benefit of our approach is that no large-deviation results are necessary for comparing tests. In a similar fashion, our approach is also more broadly applicable than the measure proposed in Hodges and Lehmann (1956), where size converges to a constant and power converges to 1, an approach which also requires large-deviation theory.

There has been other recent work in econometrics on non-local ARE measures. Kim and Perron (2009) propose using an approximate version of the Bahadur (1967) ARE 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, while maintaining an exact asymptotic comparison.

We demonstrate the broad applicability of our approach by considering several important applications of the theory. We derive an asymptotic power approximation for general use in smooth GMM problems. From there, we discuss our ARE measure and apply it to several settings. The first specific application we provide is to cluster-robust inference. The general framework we adopt is that in 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 independent observations are included in the same cluster. 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.

Our second application 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 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 third application is to linear instrumental variables (IV) models. We show that the Anderson-Rubin test (Anderson and Rubin (1949), Andrews et al. (2019)) trades off asymptotic power with the classic Wald test based on the two-stage least squares (2SLS) estimator in certain parts of the parameter space. These tests are considered equivalent under local-power comparisons, and since the Anderson-Rubin test is robust to weak instruments, generally econometricians have recommended its use.

Our distributional theory extends results in Bentkus et al. (2007), Omey and van Gulck (2009), and Shao and Zhang (2009), where one-sample t-statistics and similar types of statistics are considered. We extend the basic theory to smooth GMM problems under i.i.d. sampling, a non-smooth problem in quantile regression, and dependent data for cluster-robust inference. In Bentkus et al. (2007) these asymptotic distributions are used to motivate asymptotic power functions. We use this type of calculation to motivate our own relative efficiency comparison.

Our empirical application focuses on the case of cluster-robust inference. We use the same data set used in MacKinnon et al. (2020a). The data include 15 years of data from the American Community Survey (ACS) and corresponding minimum wage data, curated by Neumark (2019). Their application focused on testing for the correct level to cluster at, effectively providing a way to determine how to ensure that size of Wald tests is asymptotically correct. We show researchers the other side of this comparison by quantifying the power loss associated with clustering at a coarser level than necessary. Our asymptotic approximation implies that clustering at the state level, the chosen level in MacKinnon et al. (2020a), should not lead to significant power loss relative to the finer state-year level.

The rest of the paper proceeds as follows: we start by introducing the principles of our analysis in the context of a simple testing problem: hypothesis testing for means. In Section 3, we provide a treatment of the distribution of Wald statistics in GMM settings, under fixed alternatives, and provide a method for conducting power analysis. In Section 4 we discuss a new relative efficiency measure that comes out of these power calculations. In Section 5 we apply our ARE measure to cluster robust inference, quantile regression, and linear IV models and show how our analysis can inform practice in these cases. Simulations are provided in Section 6 to show the efficacy of the methods here in making finite-sample predictions. In Section 7, we apply our procedure to perform power analysis in the case of clustered sampling settings. A summary of our results is discussed in Section 8

2 A Simple Example: the Sample Mean

We begin by considering a simple testing problem: a two sided hypothesis test for the sample mean. To illustrate the basic approach, we compare the classic Wald test statistic with a cluster-robust version. Under sequences of local-alternatives, these test statistics have the same asymptotic properties, and therefore the same asymptotic power. When we compare the asymptotic distributions under fixed alternatives, we find that the asymptotic distributions differ for the two test statistics. This leads to a natural relative efficiency comparison, in which we find that when the observations are independent (i.e. both test statistics have correct asymptotic size) there is an asymptotic power loss associated with using the cluster-robust test statistic.

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] A concerned researcher suggests that we should use cluster-robust methods since the data were grouped when collected, however we know that the observations are independent and identically distributed. For all g, i, we have that $\mathbb{E}\, X_{gi} = \mu$ and $\text{Var}(X_{gi}) = \sigma^2$. Let γ and κ denote the skewness and kurtosis respectively. We would like to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. We construct Wald tests 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 suggested by another researcher. 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

[1] We can also accommodate unbalanced designs with growing cluster sizes; this type of result is also covered in Section 5.1.

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:

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

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, divided by σ^2, is asymptotically χ_1^2, and each denominator converges to σ^2 in probability. Implicitly, this effectively treats each variance estimator as equal to its probability limit. The same logic holds in the case of a sequence of local alternatives, where we consider $\mu_n = \mu_0 + \delta/\sqrt{n}$. 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 $\chi_1^2(\delta^2/\sigma^2)$.

Now, let $\mu = \Delta + \mu_0$. For discussing our results, it is useful to define the non-centrality parameter:

$$\xi := \frac{\Delta}{\sigma}$$

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 ξ^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 ξ^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:

$$\sqrt{n}\left(W_a - \xi^2\right) \Rightarrow \mathcal{N}\left(0, \xi^2\Sigma_a\right) \tag{4}$$

where

$$\Sigma_h = (\kappa - 1)\xi^2 - 4\gamma\xi + 4 \tag{5}$$

$$\Sigma_c = \Sigma_h + 2(M-1)\xi^2 \tag{6}$$

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 later relax this assumption, and doing so does not change the main conclusions. Even though our observations are i.i.d., the variance estimator in (2) involves the sum over G i.i.d. cluster-sums, whereas in the variance estimator 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: $G/n = 1/M$. The other effect is that if we expand the variance estimators in (2) and (1), the cluster-robust variance estimator will have all the same terms as the homoskedastic variance estimator, plus some additional terms. When considering the probability limit, these extra terms have mean zero and disappear. They show up in the asymptotic variance, inflating the tails of the test statistic.

We now 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 power:

$$P(nW_a > C_\alpha) \to 1 - F_{\chi^2_1(\delta^2/\sigma^2)}(C_\alpha), \quad a \in \{h, c\}, \quad \delta = \sqrt{n}(\mu_n - \mu_0) \tag{7}$$

where δ is the local parameter previously defined. This non-central chi-square distribution is the same regardless of which variance estimator we use. Thus, the asymptotic power comparisons under local alternatives do not distinguish between Wald tests where different consistent variance estimators are used; the first order asymptotics are the same for both test statistics.

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 compare the test statistics with respect to their asymptotic power properties. Note that $\Sigma_h < \Sigma_c$ as long as $M > 1$. We consider the power of the test, rearranging and normalizing the test statistic based on

the asymptotic distribution in (4):

$$\begin{aligned} P(nW_a > C_\alpha) &= P\left(W_a - \xi^2 > \frac{C_\alpha}{n} - \xi^2\right) \\ &= P\left((\xi^2\Sigma_a)^{-1/2}\sqrt{n}\left(W_a - \xi^2\right) > \frac{C_\alpha}{\sqrt{n\xi^2\Sigma_a}} - \sqrt{\frac{n\xi^2}{\Sigma_a}}\right) \end{aligned} \tag{8}$$

We cannot generally compute probabilities such as those in (8). The term on the righthand side is diverging to $-\infty$ and therefore we generally cannot assume the central limit theorem provides a good approximation here.

Part of the appeal of local power analysis is that asymptotically power is in $(0, 1)$. Under fixed alternatives, we construct a sequence of critical values C_n^a such that $P(nW_a > nC_n^a) \to 1 - \beta \in (0, 1)$, and for relative efficiency we focus on the case that $1 - \beta \in (1/2, 1)$. In this way, the sequence of critical values tells us about the speed at which the power converges to 1. We showed that $W_a \xrightarrow{P} \xi^2$. Thus, we choose a sequence C_n^a which is local to the non-centrality parameter ξ^2, and approaches this limit from below.

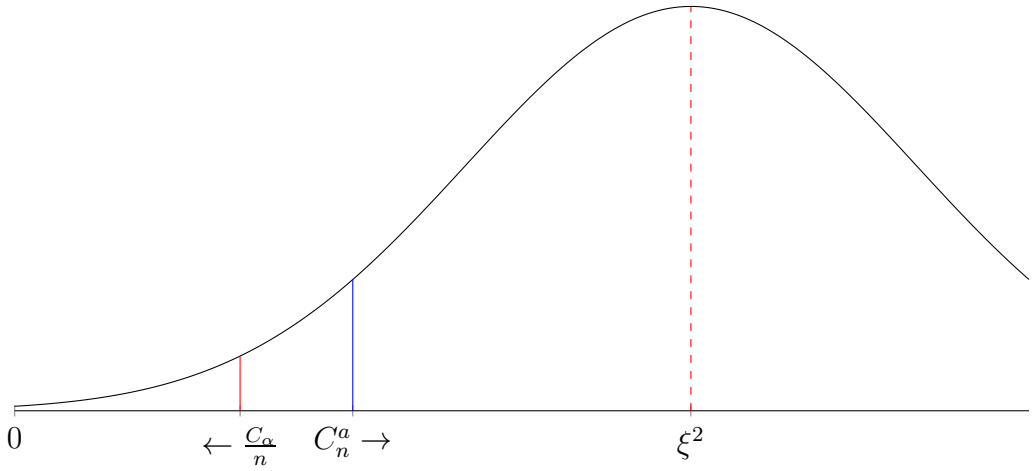

Figure 1: The distribution of W_a will concentrate around the non-centrality parameter ξ^2, while the critical value for the test C_α/n converges to 0. The chosen sequence C_n^a will converge to the non-centrality parameter at the correct rate so that the power of this sequence of tests is non-degenerate asymptotically.

In Figure 1, we lay out the relationship between the non-centrality parameter, C_n^a, and C_α/n. We want to gain insight into asymptotic behavior of the exact power in (8), and therefore choose C_n^a converging to ξ^2 from below for our relative efficiency comparison. Note that any sequence local to ξ^2 will lead to non-degenerate power. We choose the sequence of

critical values:

$$C_n^a := n\left(\xi^2 - \frac{t\Sigma_a^{1/2}}{\sqrt{n}}\right) \tag{9}$$

The following is valid for $t \in \mathbb{R}$, however the case that $t > 0$ is the relevant choice for large n, as this will imply power is above $1/2$ asymptotically:

$$P(nW_a > C_n^a) = P\left(\Sigma_a^{-1/2}\sqrt{n}\left(W_a - \xi^2\right) > -t\right) \to \Phi(t)$$

This asymptotic power approximation was proposed in Bentkus et al. (2007) in the case of the 1-sample t-test. We use it to motivate an asymptotic relative efficiency measure. For our two test statistics, W_h and W_c, which one requires a larger sequence of critical values to prevent power from converging to 1? Let us consider what happens when we use the sequence corresponding to W_h as critical values for tests using W_c. The asymptotic power of the tests becomes:

$$P(nW_c > C_n^h) = P\left(\Sigma_h^{-1/2}\sqrt{n}\left(W_c - \xi^2\right) > -t\right) \to \Phi\left(t\left(\frac{\Sigma_h}{\Sigma_c}\right)^{1/2}\right) \tag{10}$$

The last term in (10) is smaller than $\Phi(t)$ for all $t > 0$, since $\Sigma_h < \Sigma_c$. Thus, for the same sequence of critical values, the test using W_h outperforms the test using W_c.

It is instructive to compare this procedure with the local asymptotic power comparison we conducted previously. When comparing local asymptotic power, the effective non-centrality parameter $n\xi^2$ is localized around 0. This implies that asymptotically, the test statistic is on the same scale as conventional critical values. In our comparison, the critical values are localized to the effective non-centrality parameter, and analysis is conducted local to that sequence. Our relative efficiency measure can also be compared to the measure developed in Bahadur (1967). In that paper, a sequence of critical values is derived from the behavior of p-values under a fixed alternative. The rate at which that sequence disappears is then compared across test statistics in terms of how quickly the type-I error rate disappears. In this paper we specify the sequence of critical values and compare the asymptotic power of tests under the same sequence of critical values. Both procedures can be interpreted as situations where the type-I error converges to 0 and the power is asymptotically non-degenerate. A benefit of our analysis is that we only require a central limit theorem, and do not require large deviation theorems. We will revisit this point in our applications, where often we cannot compute large-deviation type probabilities.

3 Fixed-Alternative 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 show the test statistics is asymptotically normal under a fixed alternative, and then discuss how to use the approximation for power calculations.

3.1 GMM-Based Wald Statistics under Fixed-Alternatives

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

$$\mathbb{E}\, g(X_i,\, \beta) = 0 \tag{11}$$

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:

$$\sqrt{n}(\hat{\beta} - \beta) \Rightarrow \mathcal{N}(0,\, V) \tag{12}$$

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:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0 \tag{13}$$

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:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(X_i, \hat{\beta})g(X_i, \hat{\beta})', \quad \hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta'} g(X_i, \hat{\beta}), \quad \hat{V} = (\hat{Q}'\hat{\Omega}^{-1}\hat{Q})^{-1} \tag{14}$$

It is then straightforward to form Wald test statistics to test (13):

$$nW_n = \frac{n(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{V}\ell} \tag{15}$$

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:

$$nW_n \Rightarrow \chi^2_1(\delta^2/\ell' V \ell) \tag{16}$$

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 (12).

Assumption 3.1. *There is a unique β^* such that $\beta = \beta^*$ satisfies (11), $\frac{1}{n}\sum_{i=1}^n g(X_i, \hat{\beta}) = o_P(1/\sqrt{n})$, and (12) holds for $\beta = \beta^*$.*

Rather than restating standard regularity conditions, this assumption implies 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):

$$W_n - \frac{\Delta^2}{\ell' V \ell} = \frac{(\hat{\theta} - \theta)^2}{\ell' \hat{V} \ell} \tag{17}$$

$$+ \frac{2\Delta(\hat{\theta} - \theta)}{\ell' \hat{V} \ell} \tag{18}$$

$$+ \frac{\Delta^2}{\ell' \hat{V} \ell} - \frac{\Delta^2}{\ell' V \ell} \tag{19}$$

As before, we will define the probability limit of W_n as $\xi = \Delta / \sqrt{\ell' V \ell}$. The righthand side of (17) 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 (18) will be asymptotically normal. We strengthen the original assumptions to ensure asymptotic normality of (19). Expanding this term, we get two components depending on $\hat{\Omega}$ and $\hat{Q}$:

$$\begin{aligned} \frac{\Delta^2}{\ell' \hat{V} \ell} - \xi^2 = &- \frac{\xi^2}{\ell' V \ell} \operatorname{tr}\left[\Omega^{-1} Q V \ell \ell' V Q' \Omega^{-1} (\hat{\Omega} - \Omega)\right] \\ &+ \frac{2\xi^2}{\ell' V \ell} \operatorname{tr}\left[V \ell \ell' V Q' \Omega^{-1} (\hat{Q} - Q)\right] + o_P(1/\sqrt{n}) \end{aligned} \tag{20}$$

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 (14) 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 (14) 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 Σ such that:*

$$\sqrt{n}\left(W_n - \xi^2\right) \Rightarrow \mathcal{N}\left(0,\, c' \Sigma c\right) \tag{21}$$

The form of c and Σ 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, $c' \Sigma c$ corresponds to the variance of particular linear functionals of

$$a_1' g(X_i, \beta),\ a_2' g(X_i, \beta) g(X_i, \beta)' a_2,\ a_3' \frac{\partial}{\partial \beta} g(X_i, \beta) a_4 \tag{22}$$

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.

3.2 Power Calculations

We can use the central limit-theorem result in Theorem 1 to approximate power, just as we did in Section 2. Using (21) as a guide, consider the rejection probability of the standard Wald test:

$$P(nW_n > C_\alpha) = P\left(\frac{\sqrt{n}(W_n - \xi^2)}{\sqrt{c'\Sigma c}} > \frac{C_\alpha}{\sqrt{nc'\Sigma c}} - \frac{\sqrt{n}\xi^2}{\sqrt{c'\Sigma c}} \right) \tag{23}$$

The event in the righthand side of (23) involves a term obeying a central limit theorem as in (21), and a term that diverges to $-\infty$. For a sequence of critical values of the form:

$$C_n := n\left(\xi^2 - \frac{t\sqrt{c'\Sigma c}}{\sqrt{n}} \right)$$

we have that $P(nW_n > C_n) \to \Phi(t)$, for any t. Thus, our proposal is to use:

$$\hat{t}_n := \frac{\sqrt{n}\xi^2}{\sqrt{c'\Sigma c}} - \frac{C_\alpha}{\sqrt{nc'\Sigma c}} \tag{24}$$

with an asymptotic power function given as $\Phi(\hat{t}_n)$. There are two important implications that

come out of these calculations. First, this power calculation under fixed alternatives relies only on a central limit theorem. No large or moderate deviation-type results are required. Second, there is additional information present here that is not present when computing power using local asymptotics. There, asymptotic power depends on the local parameter $\delta^2/\ell' V \ell$ and the critical value C_α. Here, we supplement the non-centrality parameter ξ^2 and the critical value C_α with an additional variance term $c'\Sigma c$, which will be relevant when ξ^2 is sufficiently far from 0.

4 A New Relative Efficiency Comparison

In this section we outline how to use the previously discussed power calculations to motivate a new relative efficiency measure. Suppose we have a pair of test statistics W_n, R_n for testing a point null $H_0 : \theta = \theta_0$ about a scalar parameter $\theta \in \mathbb{R}$, such that under the null hypothesis:

$$nW_n \Rightarrow \chi^2_1$$
$$nR_n \Rightarrow \chi^2_1$$

Further, suppose that there exist sequences of constants $a_n, b_n > 0$, $a_n, b_n \to \infty$, $a_n, b_n = o(n)$ and constants $\xi, \nu \in \mathbb{R}$, $\sigma, \omega > 0$ such that under a fixed alternative $\theta \neq \theta_0$:

$$a_n(W_n - \xi^2) \Rightarrow \mathcal{N}(0, \sigma^2)$$
$$b_n(R_n - \nu^2) \Rightarrow \mathcal{N}(0, \omega^2)$$

Note that this also implies each test is consistent, so that power converges to 1. How should we go about comparing those two test statistics? From the calculations in the previous section, we can approximate the asymptotic power, however the sequences of critical values used will be different for the two test statistics, and in particular the sequences are:

$$C_n = n\left(\xi^2 - \frac{t\sigma}{a_n}\right)$$
$$D_n = n\left(\nu^2 - \frac{t\omega}{b_n}\right)$$

Specifying a particular alternative pins down the sequence of critical values in each case. Thus, if we specify an alternative θ, and then choose the same sequence of critical values for both test statistics, say C_n, under the null hypothesis the size of each sequence of tests will

converge to 0. Under the particular alternative θ we choose, we have that:

$$P(nW_n > C_n) \to \Phi(t)$$

How should we compare this with the asymptotic power of R_n? Using the same sequence of critical values for this test statistics, we have:

$$\begin{aligned} P(nR_n > C_n) &= P(nR_n > D_n + (C_n - D_n)) \\ &\to \Phi\left(t - \frac{b_n}{\omega}\left(\xi^2 - \nu^2 - t\left(\frac{\sigma}{a_n} - \frac{\omega}{b_n}\right)\right)\right) \end{aligned} \tag{25}$$

Notice that the limit (25) is valid even if the argument diverges, since we are not making any statement of the relative error in the approximation, so shared degeneracies at 0, 1/2 or 1 still imply the shared limit. Clearly, the dominant term is $b_n(\xi^2 - \nu^2)$. Thus, if $\xi^2 > \nu^2$, then (25) converges to 0. Thus, the sequence of critical values leading to non-degenerate power for W_n leads to no power asymptotically for R_n, and thus we prefer W_n to R_n. Now, suppose that $\xi^2 = \nu^2$. Then we are left with:

$$\Phi\left(t\frac{\sigma}{\omega}\frac{b_n}{a_n}\right)$$

Recall from Figure 1 that the choice of $t > 0$ seems to be more relevant for comparing the performance of tests under fixed alternatives. Thus, if $b_n/a_n \to 0$, then the asymptotic power converges to 1/2 for R_n, whereas for W_n the asymptotic power is in $(1/2, 1)$. Thus, when the convergence rate of R_n is slower than the rate for W_n, we prefer W_n. Lastly, if $b_n/a_n \to 1$, then we prefer W_n when $\sigma < \omega$.

Based on these arguments, we propose using a lexicographic preference ordering over test statistics. First, compare non-centrality parameters. The test statistic with the larger non-centrality parameter is preferred. If the non-centrality parameters are the same, as they are in different Wald tests using the same point estimator, compare the rates of convergence. If the rates of convergence are the same, then look at the asymptotic distribution of the test statistics under fixed alternatives, and choose the test statistic with the smaller asymptotic variance. In Section 5, we cover examples which illustrate how to apply this procedure in each of these cases.

5 Applications and Extensions

Evaluating the cost of cluster-robust inference involves comparing the variances (and possibly rates) of the test statistics involved. Choosing a kernel variance estimator for standard errors in a quantile regression environment requires a choice of the convergence rate (in a particular range) as well as a choice of the kernel, which affects the asymptotic variance of the test statistic. Lastly, in a linear IV environment, we compare the Anderson-Rubin test to the Wald test using the 2SLS estimator under strong identification. These two tests lead to different non-centrality parameters, and we can specify in which parts of the parameter space each test is expected to have higher power.

5.1 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 independent observations are included in the same cluster for the purpose of variance estimation. When a finer cluster level is appropriate, such as classroom, using a coarser cluster, such as school, will lead to asymptotic efficiency loss.

We consider here the just-identified case for 2SLS, where we treat OLS as a special case. For our purposes 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 we denote observation i in coarse clusters d and sub-cluster g. Implicitly, the cluster level g is nested in only one coarse cluster d. We will sometimes make this explicit notationally, and use $g(d)$ to denote that cluster g is nested in d.$Z_{g(d)}$ 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 5.1. *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:

$$a_n := (Q_n^{-1})'\ell \tag{26}$$
$$b_n := V_n\ell \tag{27}$$

We then define:

$$Y_{g(d)} := \begin{pmatrix} Z_{g'(d)}\varepsilon_{g(d)} \\ a_n'(Z_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'Z_{g(d)} - \mathbb{E}\, Z_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'Z_{g(d)})a_n \\ a_n'(Z_{g(d)}'X_{g(d)} - \mathbb{E}\, Z_{g(d)}'X_{g(d)})b_n \end{pmatrix} \tag{28}$$

$$Y_{\bullet d} := \begin{pmatrix} \sum_{g=1}^{G_d} Z_{g(d)}'\varepsilon_{g(d)} \\ a_n'\left(\left[\sum_{g=1}^{G_d} Z_{g(d)}'\varepsilon_{g(d)}\right]\left[\sum_{g=1}^{G_d} Z_{g(d)}'\varepsilon_{g(d)}\right]' - \sum_{g=1}^{G_d}\mathbb{E}\, Z_{g(d)}'\varepsilon_{g(d)}\varepsilon_{g(d)}'Z_{g(d)}\right)a_n \\ a_n'\left(\sum_{g=1}^{G_d} Z_{g(d)}'X_{g(d)} - \sum_{g=1}^{G_d}\mathbb{E}\, Z_{g(d)}'X_{g(d)}\right)b_n \end{pmatrix} \tag{29}$$

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:

$$\Xi_n^G := \frac{1}{n^2}\sum_{d=1}^{D}\sum_{g=1}^{G_d}\mathbb{E}\, Y_{g(d)}Y_{g(d)}' \tag{30}$$

$$\Xi_n^D := \frac{1}{n^2}\sum_{d=1}^{D}\mathbb{E}\, Y_{\bullet d}Y_{\bullet d}' \tag{31}$$

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 5.2.

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

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. The last assumption implies that the linear functional of interest is in fact estimable.

Assumption 5.3. *For r_G, r_D in Assumption 5.1, 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 2. *Under assumptions Assumptions 5.1-5.3, we have that there exist sequences* Ξ_n^A, ν_n, *such that for* $A \in \{G, D\}$,

$$\frac{1}{\sqrt{\nu_n' \Xi_n^A \nu_n}} \left(W_n^A - \xi_n^2\right) \Rightarrow \mathcal{N}(0, 1) \tag{32}$$

Furthermore, for all n, $\Xi_n^G \leq \Xi_n^D$, *with equality only if the cluster levels are in fact equal. Thus, asymptotically the Wald test using* W_n^G *will be more powerful than the test using* W_n^D.

$\nu_n' \Xi_n^A \nu_n$ involves both the convergence rate of each test statistic and the asymptotic variance of each test statistic, and therefore we will provide part of the argument here as to why we do not need to separate those two parts. Returning to the example of the sample mean, we perform similar calculations to (10). Consider the sequence of critical values leading to nondegenerate power for tests using W_n^G:

$$C_n^G = n\left(\xi_n^2 - t\sqrt{\nu_n' \Xi_n^G \nu_n}\right)$$

This sequence leads to asymptotic power $\Phi(t)$ for the test rejecting when $nW_n^G > C_n^G$. If instead we use W_n^D, we have:

$$\begin{aligned} P(nW_n^D > C_n^G) &= P\left((\nu_n' \Xi_n^D \nu_n)^{-1/2}(W_n^D - \xi_n^2) > -t\sqrt{\frac{\nu_n' \Xi_n^G \nu_n}{\nu_n' \Xi_n^D \nu_n}}\right) \\ &\leq P\left((\nu_n' \Xi_n^D \nu_n)^{-1/2}(W_n^D - \xi_n^2) > -t\right) \\ &\to \Phi(t) \end{aligned}$$

when $t > 0$, which we have argued is the relevant region for the relative efficiency comparisons. Thus, for asymptotic power in $(1/2, 1)$, there is a cost from using W_n^D instead of W_n^G.

It turns out that the difference $\nu_n'(\Xi_n^G - \Xi_n^D)\nu_n$ has a simple form, and this gives us some insight into when we expect these differences to be particularly stark. 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:

$$\nu_n'(\Xi_n^G - \Xi_n^D)\nu_n = \xi_n^4 \sum_{d=1}^{D} \sum_{g=1}^{G_d} \Pi_{g(d)} \left(\sum_{h \neq g} \Pi_{h(d)} \right) \tag{33}$$

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. From (33), this difference is increasing in ξ_n^2, which suggests that the tests will behave similarly when ξ_n^2 is small. This is in agreement with a local-power analysis. 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:

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

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 with homogenous cluster sizes 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 uneccessary clustering. Our results, being asymptotic in nature, also hold 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 suggests that when the fine clustering level is the correct level we can quantify the penalty for using coarser clusters. Their procedure provides researchers with information about when tests will have incorrect asymptotic size. We provide another perspective, so that researchers can evaluate and weigh both their concern for having tests with incorrect size and any loss in power from using a conservative clustering scheme.

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.

5.2 Quantile regression

In this section we apply our procedure to variance estimation in the context of quantile regression. One way to perform asymptotically valid inference in this setting involves using a kernel density estimator to estimate the asymptotic variance. We use our tools previously developed to provide insight into both the choice of a bandwidth and the choice of kernel. We will provide only a partial decision in the context of choosing the bandwidth, however we will still be able to provide researchers with some guidance. 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, therefore we need to extend our results from Section 3.1. 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:

$$y_i = x_i' \beta_\tau + \varepsilon_i, \quad Q_\varepsilon(\tau | x_i) = 0 \tag{34}$$

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 5.4. *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:

$$\hat{\Omega}_\tau = \frac{1}{n}\sum_{i=1}^{n} \tau(1-\tau) x_i x_i' \tag{35}$$

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 $\hat{\Omega}_\tau$.[2] We will need to make several assumptions on the kernel used, the conditional density of ε_i, and the bandwidth choice h_n.

Assumption 5.5. *The kernel function K is symmetric, of bounded variation, and normalized such that:*

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

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.

Assumption 5.6. *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 5.7. $h_n = o(\log n/\sqrt{n})$.

[2] 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.

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:

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

We also define the matrix A, and our non-centrality parameter ξ:

$$\xi := \frac{\Delta}{\sqrt{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell}}, \quad A = 2\xi^2 Q_\tau^{-1} \ell \ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1}$$

Under these assumptions, we have the following result:

Theorem 3. *Under Assumptions 5.5-5.7, we have that:*

$$\sqrt{n h_n} \left(W_n - \xi^2 - B_n\right) \Rightarrow \mathcal{N}\left(0,\, \mathbb{E}[(x_i' A x_i)^2 f(0|x_i) R_K]\right) \tag{36}$$

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

$$B_n = \frac{1}{2} \mathbb{E}[x_i' A x_i f''(0|x_i) h_n^2] \tag{37}$$

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 (36), the first observation we make is that we do not want to undersmooth. If we choose $h_n = o(n^{-1/5})$, then this leads to a slower rate of convergence; for our asymptotic power comparisons, eliminating the bias is not worth the slower rate of convergence. Now, for our discussion of kernel choice, we fix our bandwidth in that we assume it is of the form:

$$h_n = c R_K^{\gamma_1} n^{-\gamma_2} \tag{38}$$

This includes the bandwidth proposed by Kato (2012), and standard asymptotic integrated mean-squared error optimal bandwidths. We can use the asymptotic distribution of the variance estimator to choose the bandwidth just as was done in Kato (2012), however our result for choosing the kernel is valid under other choices of the bandwidth.

Proposition 1. *Using the Epanechnikov kernel or Gaussian kernel leads to higher asymptotic power compared with using the uniform kernel, whenever the bandwidth used is of the form in* (38)*, for* $\gamma_1 < 1$

Generally, the bandwidth choice depends on R_K so that $\gamma_1 \neq 0$, however $\gamma_1 < 1$ in all bandwidth selection rules known to us. 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.

5.3 Wald and Anderson-Rubin Under Strong Identification

In this section we consider tests with different non-centrality parameters under a fixed alternative. Up until this point, we have implicitly discussed consistent variance estimators in the sense that consistency applies to all points in the parameter space. In general, asymptotically valid tests can be developed by constructing variance estimators which are consistent under the null hypothesis, but might not be consistent more generally.

We formalize this comparison in the context of a linear instrument variables model:

$$\begin{aligned} y_i &= x_i\theta + \varepsilon_i \\ x_i &= z_i\pi + \upsilon_i \end{aligned} \tag{39}$$

We follow the exposition in Andrews et al. (2019), and presume that y_i, x_i, and z_i have already had the effects of any other control variables partialled-out. Our analysis focuses

on the case that identification is strong: there exists $C > 0$ such that $\pi^2 \geq C$. We also focus on the case $x_i, z_i \in \mathbb{R}$, and our conclusions here readily generalize to the just-identified case when $x_i, z_i \in \mathbb{R}^p$. We do not discuss over-identified settings, as there it is well known that Anderson-Rubin has some deficiencies with respect to asymptotic power under local alternatives. We will also assume that the errors are homskedastic for expositional simplicity. The two test statistics we consider for testing $H_0 : \theta = \theta_0$ under a two-sided alternative are:

$$W_n = \frac{(\hat{\theta} - \theta_0)^2 \hat{Q}^2}{\frac{1}{n}\sum_{i=1}^n (y_i - x_i\hat{\theta})^2}$$
$$R_n = \frac{(\hat{\theta} - \theta_0)^2 \hat{Q}^2}{\frac{1}{n}\sum_{i=1}^n ((y_i - z_i\hat{\pi}\hat{\theta}) - (x_i - z_i\hat{\pi})\theta_0)^2}$$

where $\hat{\theta}$ is the 2SLS estimator, $\hat{\pi}$ and $\hat{\pi}\hat{\theta}$ are the first-stage and reduced-form OLS estimators respectively, and $\hat{Q} = \frac{1}{n}\sum_{i=1}^n z_i x_i$. Note that if we replace θ_0 with $\hat{\theta}$ in the denominator of the definition of R_n, we have that $R_n = W_n$. Under the null hypothesis:

$$nW_n \Rightarrow \chi^2_1$$
$$nR_n \Rightarrow \chi^2_1$$

Let $\sigma^2 := \mathbb{E}\,\varepsilon_i^2$, $\sigma_\upsilon^2 := \mathbb{E}\,\upsilon_i^2$, $\rho := \mathbb{E}\,\varepsilon_i\upsilon_i$, and $\mathbb{E}\,z_i x_i = Q$. Then, under a fixed-alternative $\theta = \theta_0 + \Delta$, we have the following probability limits of our test statistics:

$$W_n \xrightarrow{P} \frac{\Delta^2 Q^2}{\sigma^2} := \xi_W^2$$
$$R_n \xrightarrow{P} \frac{\Delta^2 Q^2}{\sigma^2 + 2\Delta\rho + \Delta^2\sigma_\upsilon^2} := \xi_R^2$$

Our comparisons in Section 4 suggest that we should favor W_n when $\xi_W^2 > \xi_R^2$, and favor R_n when the reverse holds. By examination, we see that $\xi_W^2 > \xi_R^2$ exactly when:

$$2\Delta\rho + \Delta^2\sigma_\upsilon^2 > 0 \tag{40}$$

Thus, when $\Delta\rho > 0$, we prefer the Wald test, and when $\frac{1}{2}|\Delta|\sigma_\upsilon^2 > |\rho|$ in the case that ρ and Δ have opposite signs. Thus, in our relative efficiency comparison, we prefer Anderson-Rubin to Wald, under strong identification, in the region $\{\Delta : \Delta\rho < 0, |\Delta| < 2|\rho|/\sigma_\upsilon^2\}$. In the alternative-space, this region is a compact interval with 0 on one end and $-2\rho/\sigma_\upsilon^2$ on the other. In particular, if the first-stage is particularly noisy (large σ_υ^2) or endogeneity is weak

($|\rho|$ is small) then the region where we prefer Anderson-Rubin is small. This suggests that the robustness gained from using Anderson-Rubin with respect to weak-instruments involves a trade-off for power in part of the parameter space when identification is strong. Notice that in this case, we did not compute the asymptotic distribution. In this case, our lexicographic approach outlined in Section 4 implies that once we have different non-centrality parameters for different test statistics, we have all we need to compare power under fixed alternatives.

This is a new way of comparing the test statistics, as previously it has been noted that under local-power comparisons, in a just-identified setting R_n and W_n are asymptotically equivalent. This case is nested in our approach, in the sense that ξ_R^2 converges to ξ_W^2 as $\Delta \to 0$.

6 Simulation Evidence

In this section we evaluate the finite sample predictions made by the theory we have developed up to this point. We include simulations involving cluster-robust inference, quantile regression, and IV. All computation was done in `R` (R Core Team (2021)). The `quantreg` package (Koenker (2021)) was used for the simulations using quantile regression.

6.1 Cluster Robust Inference

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.

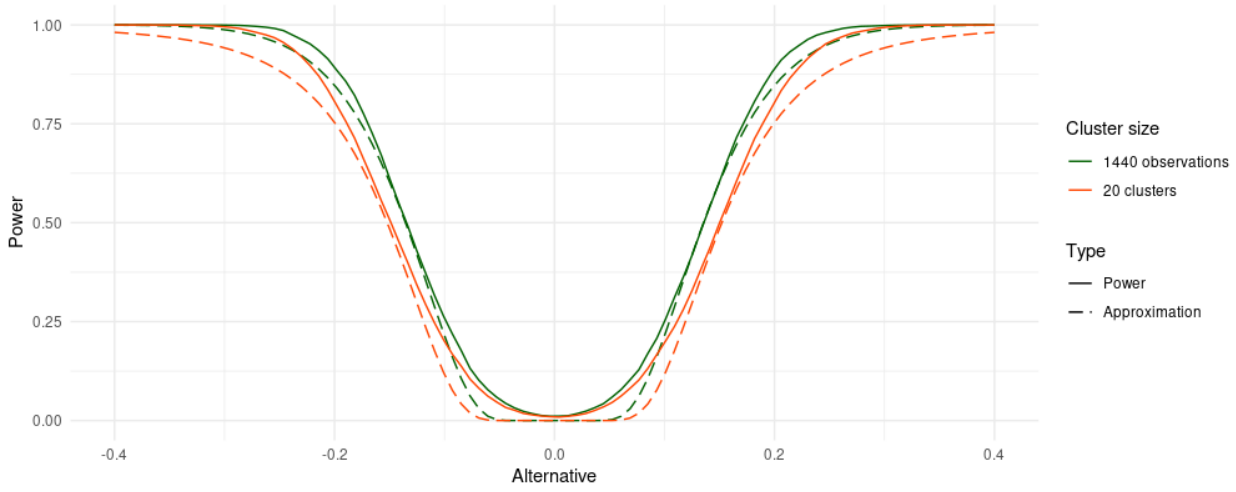

Figure 2: Sample mean, i.i.d. observations

In Figure 2, 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 solid lines are Monte-Carlo power curves. We adjusted the critical values in these simulations so that the type-I error rate is 0.01 for all tests; therefore, the tests conducted for the green solid line and the orange solid line use different critical values. The dashed lines are computed using (24), with values of C_α varying for the orange and green lines.

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.

Our theory predicts the ordering of the green and orange lines, however we do not have a result for how well the approximation based on (24) will work in practice, and therefore there is independent value in these simulations from a purely numerical perspective. We observe that the approximation does particularly poorly when the alternative is small. We would expect the local power approximation to do well in this region, and in fact it does: when we included the local power approximations in this plot (one for each cluster level, since we varied the critical values), they matched the solid curves quite well. This is due to the fact that the local power approximation based on the non-central chi-square distribution can be arrived at as an approximation to the non-central F distribution when the data are normally distributed. Overall, our approximation based on (24) does get the ordering correct throughout.

6.2 Quantile Regression

For out simulations with quantile regression, we simulated 250 i.i.d. observations based on:

$$y_i = x_i'\beta + \varepsilon_i, \quad x_i \sim t_3, \quad \varepsilon_i | x_i \sim \mathcal{N}(0,\, \|x_i\|^2)$$

We chose a simple form of heteroskedasticity that leads to the error terms having heavy tails, unconditionally. The dimension of x_i is 5, $\beta_k = 0$ for $k \neq 1$. We test the two sided hypothesis:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

We used the default settings in the `quantreg` package for computing the Bofinger (1975) and Hall and Sheather (1988) bandwidth rules. The plot was generated using 50000 repetitions.

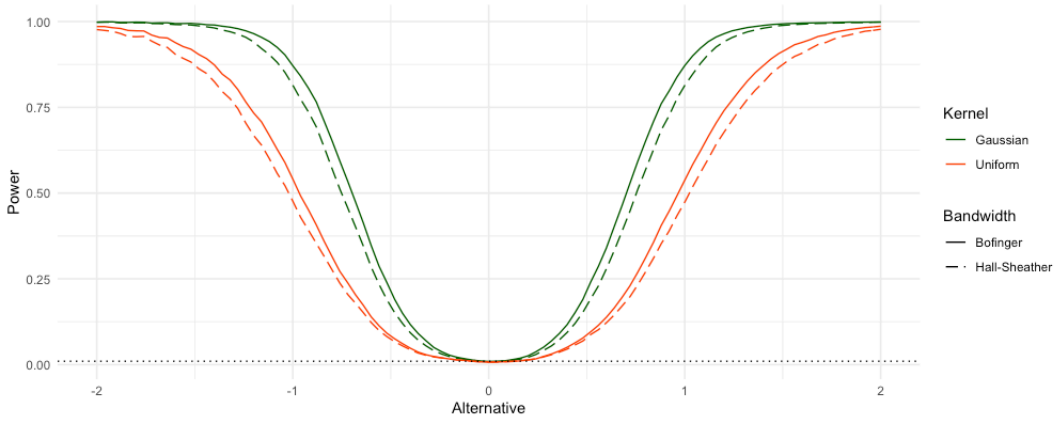

Figure 3: Quantile Regression

The solid lines are the Monte-Carlo power curves when the Bofinger (1975) bandwidth is used. The green lines result from using a Gaussian kernel, the default in `quantreg`, and the orange involve the use of the uniform kernel. Notice that our theory predicts that both use of the Hall and Sheather (1988) bandwidth rule and uniform kernel leads to less efficient inference, as seen in the plot. In finite samples, these simulations suggest that the choice of kernel is more important than the bandwidth rule. This part does run counter to the preference ordering we specified, as our theory says that if we had to choose between the Gaussian kernel with a suboptimal bandwidth and the uniform kernel with a rate optimal bandwidth, we would prefer the latter, which is clearly not reflected in Figure 3.

6.3 Anderson Rubin vs. Wald

In this final set of simulations, we consider the simplest IV model:

$$y_i = x_i\beta + \varepsilon_i$$
$$x_i = 0.2z_i + \upsilon_i$$

where x_i, $z_i \in \mathbb{R}$ and ε_i, υ_i are marginally standard normal, with correlation 0.3. We ran 50000 simulations using a sample size of 750.

In Figure 4, our theory matches quite well. The orange dashed line is the Monte-Carlo power curve for the Wald test, and the solid green line is the corresponding power curve for Anderson-Rubin. The tests use slightly different critical values so that size is controlled. On the right side of zero, we notice that Wald dominates Anderson-Rubin, as predicted by our theory, as (40) holds. To the left of zero, (40) does not hold from zero to the vertical dashes

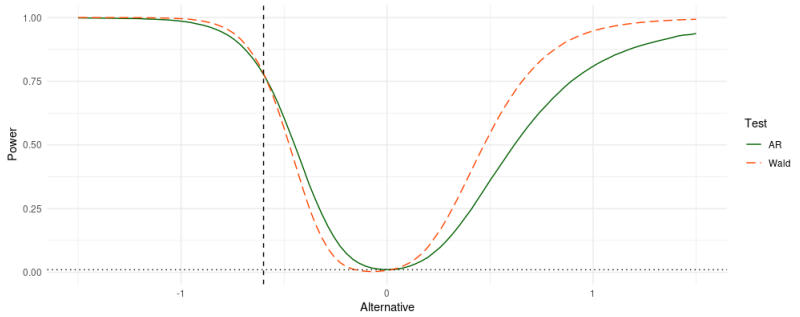

Figure 4: Anderson-Rubin vs. Wald

line, and in this region our theory states that Anderson-Rubin is preferred to Wald. To the left of the vertical dashed line, Wald is preferred once again. We see this assertions matched quite well by the simulated Monte-Carlo power curves.

7 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. The ACS data have three natural cluster levels: state-year, state, and region. We use the same data set used in MacKinnon et al. (2020a) to demonstrate how to apply their sequential testing procedure when trying to determine the appropriate clustering level.[3] 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:

$$y_{ist} = \mu + \theta \text{mw}_{st} + z_{ist}'\gamma + \delta_t \text{year}_t + \delta_s \text{state}_s + \varepsilon_{ist} \tag{41}$$

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 used comes from Neumark (2019). Details on pre-processing of the data and combining the two data sets can be found in MacKinnon et al. (2020a).

We now describe our approach to power analysis. We emphasize that our recommendations here do not replace a sequential testing procedure as proposed in MacKinnon et al.

[3]They follow several pre-processing steps that are outlined in their paper; their data can be found on the authors' websites.

(2020b). Rather, we see our methods giving researchers a more complete information set to make decisions about their testing problems. Tests determining the correct level of clustering communicate the benefit, in terms of test accuracy, of clustering at a coarser level. Our analysis communicates the cost of clustering at a coarser level when the fine level is correct. Together, the procedures give a balance of costs and benefits for researchers to make informed decisions.

Our approach is to take the approximation in (32) and use the approximating normal distribution to obtain power curves, as a function of Δ. These approximations do not control the relative error, but we argue they are still useful. To be precise, our estimated power curve is:

$$\Phi\left(\frac{\sqrt{n}\Delta^2}{(\ell'\hat{V}_n\ell)(\hat{\nu}_n'\hat{\Xi}_n\hat{\nu}_n)^{1/2}} - \frac{C_\alpha}{\sqrt{n}(\hat{\nu}_n'\hat{\Xi}_n\hat{\nu}_n)^{1/2}}\right) \tag{42}$$

In cases of practical interest, these terms will not be large. To see why, observe that $n\Delta^2/\ell'V_n\ell$ is the population version of the Wald statistic. In cases of practical interest for power comparisons, this term is smaller than 10. We need to also consider $n\nu_n'\Xi_n\nu_n$. $\nu_n'\Xi_n\nu_n$ will be small for small Δ^2, and when for sufficiently small Δ^2, the negative term will dominate, and the power will be close to 0. Thus, for Δ values of practical interest, we expect this approximation to provide good guidance.

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:

$$\sum_{g=1}^{G}\mathbb{E}(X_g'\varepsilon)^4 - \left(\mathbb{E}(X_g'\varepsilon_g)^2\right)^2 \tag{43}$$

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 5, 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 6, 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. Their results suggest that clustering at the region level should not improve size control relative to clustering at the state level, and our results imply there is a meaningful power loss from doing so. Simultaneously, their results indicate that there is a benefit to test accuracy from clustering at the state level rather than the state-year level, and our analysis indicates any loss in power if this coarser clustering is incorrect should be minimal.

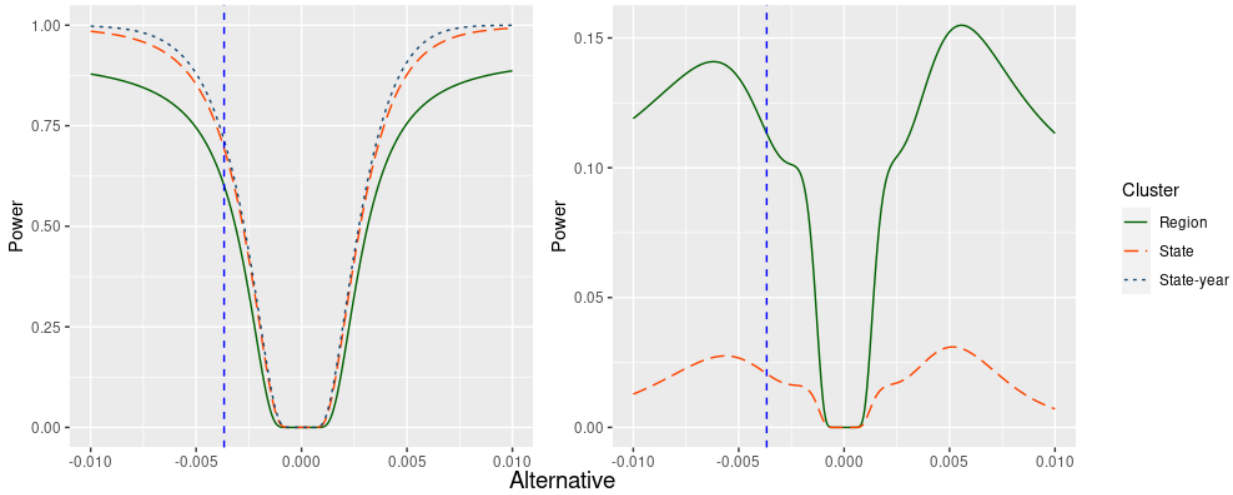

Figure 5: True clustering: state-year

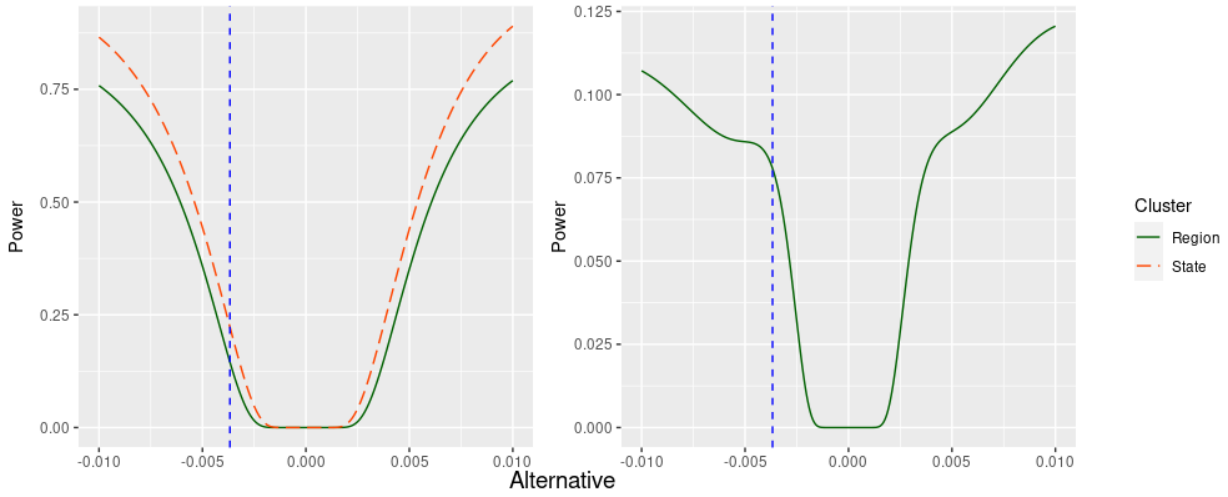

Figure 6: True clustering: state

8 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 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:

$$\sqrt{n}(\hat{\beta}-\beta) = VQ'\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\beta) + o_P(1) \tag{44}$$

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

$$\begin{aligned}
\sqrt{n}\,\mathrm{tr}(A(\hat{\Omega}-\Omega)) &= \frac{1}{\sqrt{n}}\sum_i g(X_i,\hat{\beta})'Ag(X_i,\hat{\beta}) - \sqrt{n}\bar{g}(\hat{\beta})'A\bar{g}(\hat{\beta}) - \sqrt{n}\,\mathrm{tr}(A\Omega) && (45)\\
&= \frac{1}{\sqrt{n}}\sum_i g(X_i,\beta)'Ag(X_i,\beta) - \sqrt{n}\bar{g}(\beta)'A\bar{g}(\beta) - \sqrt{n}\,\mathrm{tr}(A\Omega) && (46)\\
&\quad + \mathbb{E}[g(X_i,\beta)'A\frac{\partial}{\partial\beta'}g(X_i,\beta)]\sqrt{n}(\hat{\beta}-\beta) + o_P(1) && (47)\\
&= \frac{1}{\sqrt{n}}\sum_i g(X_i,\beta)'Ag(X_i,\beta) - \mathbb{E}[g(X_i,\beta)'Ag(X_i,\beta)] && (48)\\
&\quad + \mathbb{E}[g(X_i,\beta)'A\frac{\partial}{\partial\beta'}g(X_i,\beta)]VQ'\Omega^{-1}\frac{1}{\sqrt{n}}\sum_i g(X_i,\beta) + o_P(1) && (49)
\end{aligned}$$

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

$$\sqrt{n}\,\mathrm{tr}(B(\hat{Q}-Q)) = \frac{1}{\sqrt{n}}\sum_i \mathrm{tr}(B\frac{\partial}{\partial\beta'}g(X_i,\hat{\beta})) - \sqrt{n}\,\mathrm{tr}(B'Q) \tag{50}$$

$$= \frac{1}{\sqrt{n}}\sum_i \sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\hat{\beta})'b_k - \sqrt{n}\,\mathrm{tr}(B'Q) \tag{51}$$

$$= \frac{1}{\sqrt{n}}\sum_i \sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\beta)'b_k \tag{52}$$

$$+ \left(\sum_{k=1}^{q} \mathbb{E}\left[\frac{\partial}{\partial\beta\partial\beta'}g_k(X_i,\beta)\right] b_k\right)' \sqrt{n}(\hat{\beta}-\beta) \tag{53}$$

$$- \sqrt{n}\,\mathrm{tr}(B'Q) \tag{54}$$

$$= \frac{1}{\sqrt{n}}\sum_i \sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\beta)'b_k - \mathbb{E}\left[\sum_{k=1}^{q} \frac{\partial}{\partial\beta}g_k(X_i,\beta)'b_k\right] \tag{55}$$

$$+ \left(\sum_{k=1}^{q} \mathbb{E}\left[\frac{\partial}{\partial\beta\partial\beta'}g_k(X_i,\beta)\right] b_k\right)' \sqrt{n}(\hat{\beta}-\beta) + o_P(1) \tag{56}$$

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

$$A = \Omega^{-1}QV\ell\ell'VQ'\Omega^{-1} \tag{57}$$

$$= aa',\ a = \Omega^{-1}QV\ell \tag{58}$$

$$B = V\ell\ell'VQ'\Omega^{-1} \tag{59}$$

$$= ba',\ b = V\ell \tag{60}$$

$$c = \begin{pmatrix} \Omega^{-1}Q'V\left(\frac{-2\Delta}{\ell'V\ell}\ell - \frac{\Delta^2}{(\ell'V\ell)^2}\,\mathbb{E}[\frac{\partial}{\partial\beta'}g(X_i,\beta)Ag(X_i,\beta)] + \frac{2\Delta^2}{(\ell'V\ell)^2}\sum_{k=1}^{q}\mathbb{E}\left[\frac{\partial}{\partial\beta\partial\beta'}g_k(X_i,\beta)\right]b_k\right) \\ \frac{-\Delta^2}{(\ell'V\ell)^2} \\ \frac{2\Delta^2}{(\ell'V\ell)^2} \end{pmatrix} \tag{61}$$

and finally, Ξ is the asymptotic covariance matrix for:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} g(X_i,\beta) \\ (a'g(X_i,\beta))^2 - a'\Omega a \\ a'\frac{\partial}{\partial\beta'}g(X_i,\beta)b - a'Qb \end{pmatrix} \Rightarrow \mathcal{N}(0,\Xi) \tag{62}$$

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 (17)-(20), and applying the standard multivariate central limit theorem.

A.2 Proof of Theorem 2

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

$$
S_n := \frac{1}{n}\sum_{g=1}^{G} \begin{pmatrix} Z_g'\varepsilon_g \\ \ell' Q_n^{-1}(Z_g'\varepsilon_g\varepsilon_g'Z_g - \mathbb{E}[Z_g'\varepsilon_g\varepsilon_g'Z_g])(Q_n')^{-1}\ell \\ \ell' Q_n^{-1}(Z_g'X_g - \mathbb{E}\, Z_g'X_g)Q_n^{-1}\Omega_n(Q_n')^{-1}\ell \end{pmatrix} \tag{63}
$$

Define:

$$
c_n := \frac{1}{n}\sum_{g=1}^{G} \mathbb{E}[X_g'Z_g a_n a_n' Z_g'\varepsilon_g] \tag{64}
$$

$$
\xi_n := \Delta/\ell' V_n \ell \tag{65}
$$

$$
Y_g := \frac{1}{n} \begin{pmatrix} Z_g'\varepsilon_g \\ a_n'(Z_g'\varepsilon_g\varepsilon_g'Z_g - \mathbb{E}[Z_g'\varepsilon_g\varepsilon_g'Z_g])a_n \\ a_n'(Z_g'X_g - \mathbb{E}\, Z_g'X_g)b_n \end{pmatrix} \tag{66}
$$

The linear combination we will be interested in is:

$$
\nu_n := \begin{pmatrix} 2(\xi_n a_n - (Q_n')^{-1}c_n) \\ -\xi_n^2 \\ 2\xi_n^2 \end{pmatrix} \tag{67}
$$

First, note that by Assumption 5.2, $\Xi_n^G := \mathbb{E}\, S_n S_n'$ exists and is well-behaved, and therefore if we are going to find a limit distribution, we would expect it to be:

$$
(\nu_n'\Xi_n^G\nu_n)^{-1/2}\nu_n' S_n \Rightarrow \mathcal{N}(0,1) \tag{68}
$$

By Assumption 5.3, the Y_g are uniformly integrable, and thus Assumptions 5.1-5.2, the assumptions of Corollary 1 in Hansen and Lee (2019) are satisfied and (68) 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:

$$W_n^G - \frac{\Delta^2}{\ell' V_n \ell} \xrightarrow{P} 0 \tag{69}$$

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

$$W_n^G - \frac{\Delta^2}{\ell' V_n \ell} = \frac{(\ell'\hat{\beta} - \theta_0)^2}{\ell' \hat{V}_n^G \ell} - \frac{\Delta^2}{\ell' V_n \ell} \tag{70}$$

$$= \frac{(\ell'\hat{\beta} - \theta)^2}{\ell' \hat{V}_n^G \ell} \tag{71}$$

$$+ \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell' \hat{V}_n^G \ell} \tag{72}$$

$$+ \Delta^2 \left(\frac{1}{\ell' \hat{V}_n^G \ell} - \frac{1}{\ell' V_n \ell} \right) \tag{73}$$

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:

$$\frac{2\Delta}{\sqrt{n}\sqrt{\ell' V_n \ell}} \frac{\sqrt{n}(\ell'\hat{\beta} - \theta)}{\sqrt{\ell' V_n \ell}} + o_P(1) \tag{74}$$

This term is $O_P(1)$:

$$\frac{\sqrt{n}(\ell'\hat{\beta} - \theta)}{\sqrt{\ell' V_n \ell}} \tag{75}$$

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

$$n\ell' V_n \ell = \ell' Q_n^{-1} \sum_{g=1}^{G} \mathbb{E}\, Z_g' \varepsilon_g \varepsilon_g' Z_g (Q_n')^{-1} \ell \tag{76}$$

The rank condition on Q_n implies that this term goes to ∞, therefore (74) 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:

$$\ell'\hat{\beta} - \theta \xrightarrow{P} 0 \tag{77}$$

$$\|\hat{Q}_n - Q_n\| \xrightarrow{P} 0 \tag{78}$$

$$\|\hat{\Omega}_n - \Omega_n\| \xrightarrow{P} 0 \tag{79}$$

The rate of convergence of $\ell'\hat{\beta} - \theta$ will be the fastest rate, at least weakly. To see why, consider (74).

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

$$W_n^G - \frac{\Delta^2}{\ell' V_n \ell} = \frac{(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{V}_n^G \ell} - \frac{\Delta^2}{\ell' V_n \ell} \tag{80}$$

$$= \frac{(\ell'\hat{\beta} - \theta)^2}{\ell'\hat{V}_n^G \ell} \tag{81}$$

$$+ \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell'\hat{V}_n^G \ell} \tag{82}$$

$$+ \Delta^2 \left(\frac{1}{\ell'\hat{V}_n^G \ell} - \frac{1}{\ell' V_n \ell} \right) \tag{83}$$

(82) can be properly normalized to be asymptotically normal, so the main component of interest is (83). Using a Taylor expansion, we write:

$$\Delta^2 \left(\frac{1}{\ell'\hat{V}_n^A \ell} - \frac{1}{\ell' V_n^G \ell} \right) = -\frac{\Delta^2}{(\ell'\tilde{V}_n^G \ell)^2} \left(\ell'\hat{V}_n^G \ell - \ell' V_n \ell \right) \tag{84}$$

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 (84) into a component depending on $\hat{Q}_n$ and a component depending on $\hat{\Omega}_n$:

$$\ell'\hat{V}_n^G \ell - \ell' V_n \ell = \ell'\hat{Q}_n^{-1}\hat{\Omega}_n^G(\hat{Q}_n')^{-1}\ell - \ell' Q_n^{-1}\Omega_n(Q_n')^{-1}\ell \tag{85}$$

$$= \ell'\hat{Q}_n^{-1}\hat{\Omega}_n^G(\hat{Q}_n')^{-1}\ell - \ell'\hat{Q}_n^{-1}\Omega_n(\hat{Q}_n')^{-1}\ell \tag{86}$$

$$+ \ell'\hat{Q}_n^{-1}\Omega_n(\hat{Q}_n')^{-1}\ell - \ell' Q_n^{-1}\Omega_n(Q_n')^{-1}\ell \tag{87}$$

First, consider (86). 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:

$$\check{\Omega}_n^G := \frac{1}{n}\sum_{g=1}^{G} Z_g'\varepsilon_g\varepsilon_g' Z_g \tag{88}$$

$$\hat{\Omega}_n^G = \check{\Omega}_n^G - \frac{1}{n}\sum_{g=1}^{G} Z_g'\varepsilon_g(\hat{\beta}-\beta)'X_g'Z_g - \frac{1}{n}\sum_{g=1}^{G} Z_g'X_g(\hat{\beta}-\beta)\varepsilon_g' Z_g \tag{89}$$

$$+ \frac{1}{n}\sum_{g=1}^{G} Z_g'X_g(\hat{\beta}-\beta)(\hat{\beta}-\beta)'X_g'Z_g \tag{90}$$

We will actually show that we only need to consider (89), and (90) will be negligible, since it is of lower-order. With our moment conditions, we have that there exists C such that:

$$\left\| \frac{1}{n}\sum_{g=1}^{G} Z_g'X_g(\hat{\beta}-\beta)(\hat{\beta}-\beta)'X_g'Z_g \right\| \leq C\|\hat{\beta}-\beta\|^2 \tag{91}$$

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

$$\left\| \frac{1}{n}\sum_{g=1}^{G} Z_g'\varepsilon_g(\hat{\beta}-\beta)'X_g'Z_g \right\| \leq C\|\hat{\beta}-\beta\| \tag{92}$$

We proceed by next carefully performing a Taylor expansion of (87).

$$\ell'\hat{Q}_n^{-1}\Omega_n(\hat{Q}_n')^{-1}\ell - \ell'Q_n^{-1}\Omega_n(Q_n')^{-1}\ell = \operatorname{tr}(-2(\tilde{Q}_n)^{-1}\Omega_n(\tilde{Q}_n')^{-1}\ell\ell'(\tilde{Q}_n)^{-1}(\hat{Q}_n - Q_n) \tag{93}$$

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 (82), (86), and (93), and the constants in (84), we have:

$$W_n^G - \frac{\Delta^2}{\ell' V_n^G \ell} = \frac{2\Delta(\ell'\hat{\beta}-\theta)}{\ell'\hat{V}_n\ell} \tag{94}$$

$$- \frac{\Delta^2}{(\ell'\tilde{V}_n^G\ell)^2}\operatorname{tr}((\hat{Q}_n')^{-1}\ell\ell'\hat{Q}_n^{-1}(\hat{\Omega}_n^G - \Omega_n)) \tag{95}$$

$$+ \frac{2\Delta^2}{(\ell'\tilde{V}_n^G\ell)^2}\operatorname{tr}((\tilde{Q}_n)^{-1}\Omega_n(\tilde{Q}_n')^{-1}\ell\ell'(\tilde{Q}_n)^{-1}(\hat{Q}_n - Q_n) \tag{96}$$

$$+ O_P(1/n) \tag{97}$$

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

$$\bar{W}_n^G - \frac{\Delta^2}{\ell' V_n \ell} := \frac{2\Delta(\ell'\hat{\beta} - \theta)}{\ell' V_n \ell} - 2\frac{1}{n}\sum_{g=1}^{G} \mathbb{E}[\varepsilon_g' Z_g (Q_n')^{-1} \ell\ell' Q_n^{-1} Z_g' X_g](\hat{\beta} - \beta) \tag{98}$$

$$- \frac{\Delta^2}{(\ell' V_n \ell)^2} \operatorname{tr}((Q_n')^{-1}\ell\ell' Q_n^{-1}(\check{\Omega}_n^G - \Omega_n)) \tag{99}$$

$$+ \frac{2\Delta^2}{(\ell' V_n \ell)^2} \operatorname{tr}((Q_n)^{-1}\Omega_n (Q_n')^{-1}\ell\ell' (Q_n)^{-1}(\hat{Q}_n - Q_n) \tag{100}$$

since

$$\|W_n^G - \bar{W}_n^G\| = o_P\left(\max\left\{\|\hat{\beta} - \beta\|, \|\hat{Q}_n - Q_n\|, \|\check{\Omega}_n - \Omega_n\|\right\}\right) \tag{101}$$

Furthermore, we also have that:

$$\|\bar{W}_n^G - \Delta^2/\ell' V_n \ell - \nu_n' S_n\| = o_P\left(\max\left\{\|\hat{\beta} - \beta\|, \|\hat{Q}_n - Q_n\|, \|\check{\Omega}_n - \Omega_n\|\right\}\right) \tag{102}$$

Note that $(\|\hat{\beta} - \beta\|, \|\hat{Q}_n - Q_n\|, \|\check{\Omega}_n - \Omega_n\|)' = O_P((\nu_n' \Xi_n^G \nu_n)^{1/2})$. Thus, we have that:

$$(\nu_n' \Xi_n^G \nu_n)^{-1/2}\left(W_n^G - \frac{\Delta^2}{\ell' V_n \ell}\right) = (\nu_n' \Xi_n^G \nu_n)^{-1/2} \nu_n' S_n + o_P(1) \tag{103}$$

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 $(Y_g)_{q+1,q+1}$

$$[\Xi_n^D]_{q+1,q+1} - [\Xi_n^G]_{q+1,q+1} = \frac{1}{n^2}\sum_{d=1}^{D}\sum_{g\neq h} \mathbb{E}(a_n' Z_{g(d)}' \varepsilon_{g(d)})^2 \, \mathbb{E}(a_n' Z_{h(d)}' \varepsilon_{h(d)})^2 \geq 0 \tag{104}$$

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:

$$Y_d := \sum_{g=1}^{G_d} a_n' Z_g' \varepsilon_g \tag{105}$$

$$= \sum_{g=1}^{G_d} Y_{dg} \tag{106}$$

$$\mathbb{E}\, Y_d = \sum_{g=1}^{G_d} \mathbb{E}\, Y_{g(d)} \tag{107}$$

$$= 0 \tag{108}$$

$$\operatorname{Var}(Y_d) = \mathbb{E}\, Y_d^2 = \sum_{g=1}^{G_d} \mathbb{E}(a_n' Z_g' \varepsilon)^2 \tag{109}$$

$$= \sum_{g=1}^{G_d} \operatorname{Var}(Y_{dg}) \tag{110}$$

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

$$\mathbb{E}\, Y_d^4 = k_4(Y_d) + 3(\mathbb{E}\, Y_d^2)^2 = k_4(Y_d) + 3\left(\sum_{g=1}^{G_d} \operatorname{Var}(Y_{g(d)})\right)^2 \tag{111}$$

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:

$$[\Xi_n^D]_{q+1,q+1} - [\Xi_n^G]_{q+1,q+1} = \frac{1}{n^2}\sum_{d=1}^{D}\sum_{g\neq h}\mathbb{E}(Y_{g(d)})^2\,\mathbb{E}(Y_{h(d)})^2 \tag{112}$$

$$= \frac{1}{n^2}\sum_{d=1}^{D}\sum_{g\neq h}\mathbb{E}(a_n' Z_{g(d)}'\varepsilon_{g(d)})^2\,\mathbb{E}(a_n' Z_{h(d)}'\varepsilon_{h(d)})^2 \geq 0 \tag{113}$$

$$\tag{114}$$

Now, consider the critical values:

$$C_n = n\left(\frac{\Delta^2}{\ell' V_n^G \ell} - t(\nu_n'\Xi_n^G\nu_n)^{1/2}\right) \tag{115}$$

Under this sequence of critical values, our power approximation for using W_n^G is:

$$\begin{aligned} P(nW_n^G > C_n) &= P\left((\nu_n'\Xi_n^G\nu_n)^{-1/2}\left(W_n^G - \frac{\Delta^2}{\ell' V_n^G\ell}\right) > -t\right) \\ &\to \Phi(t) \end{aligned}$$

When using W_n^D, we have:

$$\begin{aligned} P(nW_n^D > C_n^G) &= P\left((\nu_n'\Xi_n^D\nu_n)^{-1/2}(W_n^D - \xi_n^2) > -t\sqrt{\frac{\nu_n'\Xi_n^G\nu_n}{\nu_n'\Xi_n^D\nu_n}}\right) \\ &\leq P\left((\nu_n'\Xi_n^D\nu_n)^{-1/2}(W_n^D - \xi_n^2) > -t\right) \\ &\to \Phi(t) \end{aligned}$$

for $t > 0$.

A.3 Proof of Theorem 3

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

$$W_n - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega Q_\tau^{-1} \ell} \tag{116}$$

$$= \frac{(\ell' \hat{\beta}(\tau) - \theta)^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega} \hat{Q}_\tau^{-1} \ell} \tag{117}$$

$$+ \frac{2\Delta(\ell' \hat{\beta}(\tau) - \theta)}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega} \hat{Q}_\tau^{-1} \ell} \tag{118}$$

$$+ \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega} \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega Q_\tau^{-1} \ell} \tag{119}$$

By Slutksy's theorem, Assumption 5.4 and the bandwidth condition in Assumption 5.7, (117) is $O_P(1/n)$ and (118) is $O_P(1/\sqrt{n})$. Turning to (119), consider the expansion:

$$\frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell} \tag{120}$$

$$= \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \Omega_\tau \hat{Q}_\tau^{-1} \ell} + \frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell} \tag{121}$$

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:

$$\frac{\Delta^2}{\ell' \hat{Q}_\tau^{-1} \hat{\Omega}_\tau \hat{Q}_\tau^{-1} \ell} - \frac{\Delta^2}{\ell' Q_\tau^{-1} \Omega_\tau Q_\tau^{-1} \ell} = \operatorname{tr}(A(\hat{Q}_\tau - Q_\tau)) + o_P(1/\sqrt{n h_n}) \tag{122}$$

$$A := \frac{2Q^{-1}\Delta^2 \ell \ell' Q^{-1} \Omega Q^{-1}}{(\ell' Q^{-1} \Omega Q^{-1} \ell)^2} \tag{123}$$

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:

$$\sqrt{\frac{n}{h_n}}\left(\mathbb{P}_n\,\psi_{\hat{\beta}} - \mathbb{E}\,\psi_\beta\right) = h_n^{-1/2}\left(\mathbb{G}_n\,\psi_b|_{b=\hat{\beta}} - \mathbb{G}_n\,\psi_\beta\right) \tag{124}$$

$$+\, h_n^{-1/2}\,\mathbb{G}_n\,\psi_\beta \tag{125}$$

$$+\, \sqrt{nh_n}\left(h_n^{-1}\,\mathbb{E}\,\psi_b(z_i)|_{b=\hat{\beta}} - h_n^{-1}\,\mathbb{E}\,\psi_\beta(z_i)\right) \tag{126}$$

(126) can be bounded since:

$$\mathbb{E}\,h_n^{-1}\psi_b(z_i) = \mathbb{E}\left[x_i'Ax_i(f(x_i'(\beta_\tau - b)|x_i) + \frac{1}{2}f''(x_i'(\beta_\tau - b)|x_i)h_n^2 + o(h_n^2))\right] \tag{127}$$

Thus, by Assumption 5.4 and 5.6, (126) is $o_P(1)$. Thus, we can turn our attention to (124), since (125) will satisfy a standard Lindeberg CLT for kernel density estimators. An implication of Assumption 5.5 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,

$$\left|K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right| \leq \int \left|\mathbb{1}_{[\frac{\varepsilon_i - x_i'\delta_s}{h_n},\, \frac{\varepsilon_i - x_i'\delta_t}{h_n}]}(x)\right| d(K_1(x) + K_2(x)) \tag{128}$$

Thus, using (128), we can use Hölder's inequality to bound the mean-squared difference:

$$\mathbb{E}\left[\left(K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right)^2 \middle| x_i\right] \leq \int \mathbb{E}\left|\mathbb{1}_{[\frac{\varepsilon_i - x_i'\delta_s}{h_n}, \frac{\varepsilon_i - x_i'\delta_t}{h_n}]}(x)\right| d(K_1(x) + K_2(x))|K|_v \tag{129}$$

$$= \int \left|\int_{x_i'\delta_s + h_n x}^{x_i'\delta_t + h_n x} f(\varepsilon|x_i) d\varepsilon\right| d(K_1(x) + K_2(x))|K|_v \tag{130}$$

$$\leq \|f(\cdot|x_i)\|_\infty |K|_v^2 \|x_i\|_2 \|t - s\|_2 \tag{131}$$

Now, by Assumption 5.6:

$$\mathbb{E}\left[\left(K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right)^2\right] = O(\|t - s\|_2) \tag{132}$$

Now, we return to (124). 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-type.*

The arguments are the same as in Giné and Nickl (2016), with the finite-dimensional vector space having dimension $p + 2$, so we omit the proof.

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 (132). This means that when $h_n = o(\log n/\sqrt{n})$, (124) converges to zero in probability.

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

$$\sqrt{nh_n}\left(W_n - \frac{\Delta^2}{\ell' Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\ell} - \frac{1}{2}\,\mathbb{E}[x_i' A x_i f''(0|x_i)h_n^2]\right) \Rightarrow \mathcal{N}\left(0,\, \mathbb{E}[(x_i' A x_i)^2 f(0|x_i) R_K]\right) \tag{133}$$

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.

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.

Bofinger, E. (1975): "Optimal Condensation of Distributions and Optimal Spacing of Order Statistics," *Journal of the american statistical association*, 70, 151–154.

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.

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.

Hall, P. and S. J. Sheather (1988): "On the Distribution of a Studentized Quantile," *Journal of the Royal Statistical Society: Series B (Methodological)*, 50, 381–391.

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.

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

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. (2021): *Quantreg: Quantile Regression*.

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

MacKinnon, J. G., M. Nielsen, and M. D. Webb (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.

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.

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. (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.

R Core Team (2021): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

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

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