Alternative Asymptotics and Fixed-Alternative Power Analysis *

Samuel P. Engle†

November 17, 2023

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 under local power approximations. This paper develops a simple asymptotic framework under fixed alternatives, which leads to new conclusions. In particular, we identify that variance estimation will have a first-order impact on the efficiency of Wald tests when size tends to zero with sample size while effect sizes are fixed. We apply this framework to several applications, including cluster-robust inference and quantile regression. In the case of cluster-robust inference, we provide for an asymptotic framework in which choosing the wrong cluster size can lead to lower power of tests. Simulations demonstrate that the results are applicable to moderate sample sizes and that the results can provide a useful approximation for size in standard ranges.

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. Ignoring any efficiency implications of variance estimation is at odds with the lived experience of empirical researchers, in that robust variance estimators are often considered "conservative." In this paper we provide a first-order asymptotic theory which characterises scenarios in which variance estimation affects the asymptotic power of tests. The resulting asymptotic theory provides a theoretical foundation for several common "folk" theorems in applied work.

The choice of variance estimator is an ever 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

*I am grateful for the encouragement and advice I have received from Jack Porter, Bruce Hansen, Mikkel Sølvsten, and Harold Chiang. I also thank Xiaoxia Shi, Giuseppe Cavaliere, Annie Lee, John Stromme, Anna Trubnikova, Anson Zhou. All remaining errors are my own.

†University of Exeter Business School, Rennes Drive, Exeter, Devon, EX4 4PU, UK. s.p.engle@exeter.ac.uk

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 provide a general theory of the asymptotic behaviour of Wald test statistics under a fixed alternative. This theory allows us to answer an important question: how does the choice of variance estimator affect the behaviour of test statistics under the alternative? We find that under a sequence of tests where size tends to zero and power converges to a constant, different variance estimators lead to different behaviour. The limiting power along the sequence of tests depends not only on consistency of the variance estimator, but also the variance and possible covariance with the estimator of the parameter of interest. Second, this implies a way of conducting power analysis in a small-size, fixed-alternative regime. Simulations provide supporting evidence that this approximation is accurate in parts of the parameter space where power is high. Last, we apply this approach to several applications, including smooth generalized method of moments (GMM) problems, linear models with possible cluster dependence, and quantile regression.

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. Under fixed alternatives, we can view a re-scaled version of the tests statistic as an estimator of a non-centrality parameter: the ratio of the effect size to the variance. With this perspective in mind, the asymptotic behaviour of the test statistic depends on both the estimation of the numerator (depending on the parameter estimator) and the denominator (depending on the variance estimator). 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. It turns out that variance estimation has an impact on asymptotic power in testing regimes where size converges to 0 and power converges to some fixed constant. These asymptotics have a similar flavour to the asymptotic experiment proposed in Bahadur (1967). A benefit of our approach is that no large-deviation results are necessary. 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.

We are not the first to use asymptotics to assess the behaviour of different variance estimators. Sun, Phillips, and Jin (2008) compares fixed-bandwidth to small bandwidth asymptotics in constructing standard errors for time-series regressions. They note that traditional optimal choices of the bandwidth are chosen based on estimation criteria, rather than testing criteria, and develop a theory where the power and size of tests depend explicitly on bandwidth choice in a variety of asymptotic frameworks. Iacone, Leybourne, and Taylor (2013) uses fixed-b asymptotics to choose between tuning parameters for variance estimators and test statistics when testing for breaks in a linear trend. This paper links the estimation of the variance to the power of tests along a different sequence of tests where size tends to zero. In Kato (2012), the asymptotic distribution is derived for the kernel density estimator for the asymptotic variance in the quantile regression setting of Koenker and Bassett (1978), for the particular choice of a uniform kernel. We extend the results in Kato (2012), and link the form of the asymptotic bias and variance to power calculations.

There has also been other work in econometrics on ways of characterising the behaviour of tests under the alternative that also differ from the local asymptotic normality approach of Le Cam (2012). Kim and Perron (2009) propose using an approximate version of the Bahadur (1967) asymptotic relative efficiency measure

when comparing tests for structural breaks in time series. Canay and Otsu (2012) used Hodges-Lehmann asymptotic relative efficiency to assess the efficiency of 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 approximation of Wald test statistics for general use in GMM problems. The main result is quite general, allowing for cluster-dependence and non-smooth moment conditions. The first specific application we provide is to cluster-robust inference in the linear model. The general framework for cluster-robust inference we adopt is that in Hansen and Lee (2019). An alternative approach, the design-based approach, is discussed in Abadie, Athey, Imbens, and Wooldridge (2020). Popularized in Bertrand, Duflo, and Mullainathan (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, Athey, Imbens, and Wooldridge (2023) presents a design-based approach to choosing the appropriate cluster level, along with some finite-sample results. MacKinnon, Nielsen, and Webb (2023) provide a sequential testing procedure to detect the correct clustering level, within the model-based framework we adopt. 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 specific application is to quantile regression. 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). The default choices in the `quantreg` package in `R` and the `qreg` function in Stata are the Gaussian and Epanechnikov kernel, respectively. We show that the choice of kernel and bandwidth do have an impact on the first-order behaviour of the test statistics in our setting.

Our distributional theory extends results in Bentkus, Jing, Shao, and Zhou (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 GMM problems under sampling with cluster-dependence, including non-smooth problems such as quantile regression. 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.

The rest of the paper proceeds as follows: we start by introducing the principles of our analysis in Section 2, 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. A strategy for conduction power analysis and some simulations are provided in Section 4 to show the efficacy of the methods here in making finite-sample predictions. A summary of our results is discussed in Section 5. Throughout, Φ is used to denote the standard normal cumulative distribution function and $\Rightarrow$ is used to denote convergence in distribution.

2 An Illustrative 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 function. Hence, it is not possible to use asymptotic methods to compare such tests. To overcome this, we compare the asymptotic distributions under fixed alternatives. In this asymptotic setting 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 equal-sized groups, each containing n_g observations, for a total of $Gn_g = n$ 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:

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

We compare the test statistic we prefer, the classic Wald test-statistic, to a cluster robust version suggested by another researcher.[2] The classic Wald test statistic, assuming homskedasticity, is given by:

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

The cluster-robust Wald statistic is:

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

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

$$W_h \Rightarrow \chi_1^2$$
$$W_c \Rightarrow \chi_1^2$$

by 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 $\chi_1^2(\delta^2/\sigma^2)$.

[1] We can also accommodate unbalanced designs with growing cluster sizes; this type of result is also covered in Section 3.

[2] For simplicity, we do not include any degrees-of-freedom correction, whether for the variance estimator or number of clusters. Since we use the large-G asymptotics of Hansen and Lee (2019), these degrees of freedom corrections disappear in the limit.

Now consider the fixed alternative $\mu = \Delta + \mu_0$ and define the non-centrality parameter:

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

For $a \in \{h, c\}$, expansions of the test statistics under a fixed alternative are:

$$\frac{1}{n} W_a = \frac{(\bar{X}_n - \mu)^2}{\hat{\sigma}_a^2} + \frac{2\Delta(\bar{X}_n - \mu)}{\hat{\sigma}_a^2} + \frac{\Delta^2}{\hat{\sigma}_a^2} \tag{3}$$

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:

$$\sqrt{n}\left(\frac{1}{n} W_a - \xi^2\right) \Rightarrow \mathcal{N}\left(0,\, \Sigma_a\right) \tag{4}$$

where

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

$$\Sigma_c = \xi^2 \left(\Sigma_h + 2(n_g - 1)\xi^2\right). \tag{6}$$

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: $G/n = 1/n_g$. 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_1^2 random variable. Local alternatives give a (local) asymptotic approximation to power:

$$P(nW_a > C_\alpha) \to 1 - F_{\chi_1^2(\delta^2/\sigma^2)}(C_\alpha), \quad a \in \{h, c\}, \quad \delta = \sqrt{n}(\mu_n - \mu_0) \tag{7}$$

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 $n_g > 1$ and $\xi \neq 0$. We consider the power of the

test, rearranging and normalising the test statistic based on the asymptotic distribution in (4):

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

Part of the appeal of local power analysis is that asymptotically power is in $(0, 1)$. For this to hold true in this case, we must choose C_n^a appropriately. Under fixed alternatives, we construct a sequence of critical values C_n^a such that $P(nW_a > C_n^a) \to 1 - \beta \in (0, 1)$. In this way, the sequence of critical values tells us about the speed at which the power converges to 1. We choose a sequence C_n^a which is local to the (scaled) non-centrality parameter $n\xi^2$, and approaches this limit from below. Notice that if we set

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

then $P(nW_a > C_n^a) \to \Phi(t)$, where Φ denotes the standard normal cumulative distribution function.

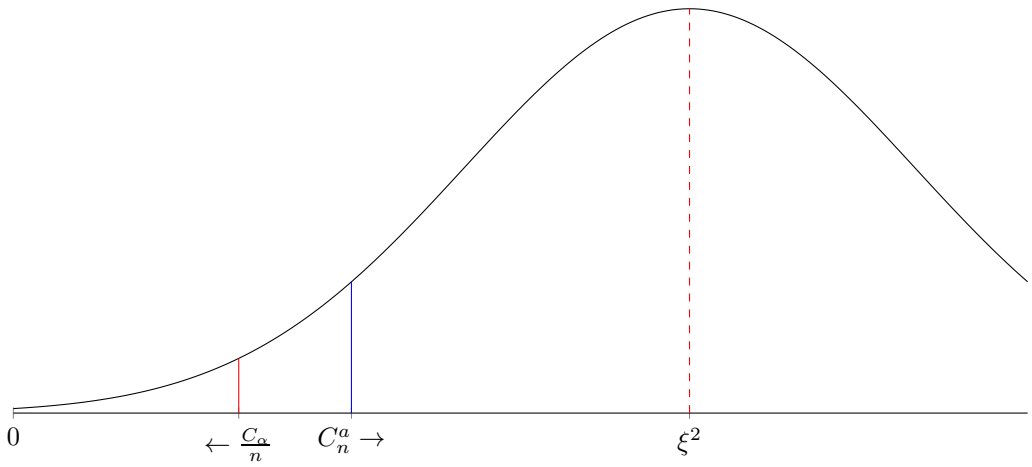

Figure 1: The distribution of $\frac{1}{n}W_a$ will concentrate around the non-centrality parameter ξ^2, while the critical value for the test C_α/n converges to 0. The chosen sequence C_n^a will converge to the non-centrality parameter at the correct rate so that the power of this sequence of tests 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 in this way will lead to non-degenerate asymptotic power, and we chose this particular sequence for convenience.

A similar asymptotic power approximation was proposed in Bentkus et al. (2007) in the case of the 1-sample t-test. In that paper they were primarily concerned with variations in behaviour of the t-test under a fixed alternative when the observations had possibly fewer than 4 finite moments. 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 has different power properties than 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 approach 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 then 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 Wald test statistics under fixed alternatives. In this section we introduce the general setup for deriving the asymptotic distribution of Wald test statistics under fixed alternatives for asymptotically linear estimators. We first give sufficient conditions for when the test statistics are asymptotically normal under a fixed alternative, and then use this first-order approximation to motivate an alternative approach to power calculations.

3.1 Test Statistics under Fixed-Alternatives

We focus our attention on parameters which are just-identified by estimating equations:

$$\mathbb{E}\,\psi(X_i, \beta) = 0 \tag{1}$$

where $X_i \in \mathcal{X}$, $\beta \in \mathcal{B} \subset \mathbb{R}^p$, and $\psi : \mathcal{X} \times \mathcal{B} \to \mathbb{R}^p$. Our focus will be on tests of a linear hypothesis. Define the parameter $\theta := L'\beta$ for a fixed matrix L of rank k. Tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ are often performed using the Wald statistic:

$$W_n := n(L'\hat{\beta} - \theta_0)'(L'\hat{V}_n L)^{-1}(L'\hat{\beta} - \theta_0) \tag{2}$$

where $\hat{V}_n$ is a suitable variance estimator of a constant matrix V_n such that $V_n^{-1/2}\sqrt{n}(\hat{\beta} - \beta) \Rightarrow \mathcal{N}(0, I_p)$. Under standard regularity conditions, W_n is asymptotically χ^2_k.

When $\theta \neq \theta_0$, W_n will diverge off to ∞, in the sense that for all $C > 0$, $P(W_n < C) \to 0$. This is clearly not useful for characterising the behaviour of W_n in large samples, and therefore we will consider suitable

centring and (re)scaling. Our first observation towards this goal is that under the same conditions providing for W_n to be asymptotically χ^2_k, for $\Delta = \theta - \theta_0$, we have that:

$$(L'\hat{\beta} - \theta_0)'(L'\hat{V}_n L)^{-1}(L'\hat{\beta} - \theta_0) - \Delta'(L'V_n L)^{-1}\Delta \xrightarrow{P} 0 \tag{3}$$

Thus, under a fixed alternative, $\frac{1}{n}W_n$ is a consistent estimator of the noncentrality parameter $\xi_n^2 := \Delta'(L'V_nL)^{-1}\Delta$. In this light, $\frac{1}{n}W_n$ is a quadratic function of $\hat{\beta}$ and a nonlinear function of $\hat{V}_n$. V_n is further decomposed into two parts, Q_n and Ω_n, so that $V_n = (Q_n'\Omega^{-1}Q_n)^{-1}$, where we will assume we have consistent estimators $\hat{Q}_n$ and $\hat{\Omega}_n$ of Q_n and Ω_n respectively. We will assume $\hat{V}_n$ is constructed as $(\hat{Q}_n'\hat{\Omega}_n^{-1}\hat{Q}_n)^{-1}$, and we consider the stochastic expansion of the difference in (3): for some non-random sequences of vectors a_n and matrices A_{1n} and A_{2n}, we have that

$$\frac{1}{n}W_n - \xi_n^2 = a_n'(\hat{\beta} - \beta) + \operatorname{tr}(A_{1n}(\hat{Q}_n - Q_n)) + \operatorname{tr}(A_{2n}(\hat{\Omega}_n - \Omega_n)) + R_n. \tag{4}$$

Just as we did in Section 2.1, we write the test statistic as a term which depends on the estimator, $a_n'(\hat{\beta}-\beta)$, and terms that depend on the components of the variance estimator, $\operatorname{tr}(A_{1n}(\hat{Q}_n-Q_n))$ and $\operatorname{tr}(A_{2n}(\hat{\Omega}_n-\Omega_n))$. Under suitable conditions, the remainder R_n is over lower order than the other terms and we will be able to specify a convergence rate and limiting distribution for $\frac{1}{n}W_n$ under the fixed alternative $\theta = \theta_0 + \Delta$. In particular, we will provide conditions such that there exist sequences Ξ_n and B_n such that

$$\Xi_n^{-1/2}\left(\frac{1}{n}W_n - \xi_n^2 - B_n\right) \Rightarrow \mathcal{N}(0,1) \tag{5}$$

We now more formally lay out the environment we are interested in, along with the assumptions we will make that assure existence of such sequences. We will point out the content and applicability of the assumptions in a few examples motivated by particular econometric applications.

Assumption 1. *(Cluster dependence) We will allow for cluster dependence, and in particular we will denote the observations X_{ig} for observation i in cluster g. Observations are independent across clusters, and we allow for arbitrary cluster dependence within cluster g, where there are G clusters. If cluster g has n_g observations, the overall sample size is $n = \sum_{g=1}^G n_g$. We require $G \to \infty$ as $n \to \infty$, as for some $r \in [2, \infty)$ and $C < \infty$, we have that*

$$\frac{\left(\sum_{g=1}^G n_g^{2r}\right)^{2/r}}{n} \leq C, \quad \lim_{n\to\infty} \max_g \frac{n_g^4}{n} = 0. \tag{6}$$

This assumption is identified in Hansen and Lee (2019) as ensuring that clusters cannot be too heterogeneous or grow too quickly with n. Notice that we use the version in Assumption 3 from that paper, as we will be characterising the asymptotic behaviour of estimators of second moments. In this case, the estimator of the variance cannot be written as a sum over i, but is rather written as a sum over the clusters g, which requires the second part of (6) to have n_g^4 instead of n_g^2. The constant r will be connected to the number of finite moments required by the influence functions below: higher cluster-size heterogeneity will require more bounded moments. This all provides for application of a Lindeberg-Lévy CLT to functions of the observations at the cluster-level, and serve as sufficient conditions for the influence of any individual cluster to be uniformly asymptotically negligible. It trivially applies for large enough samples where the cluster size n_g is constant, including $n_g = 1$, but also applies when the clusters are growing, e.g. $n_g = n^\alpha$ for some $\alpha < (r-2)/2(r-1)$, or a mix of the two, where only some clusters are growing. For more discussion of these

examples and other considerations, see Hansen and Lee (2019).

The next assumption confirms the definition of β, $\hat{\beta}$, and provides for W_n leading to valid inference when $\beta = \beta_0$. Before doing this, it is helpful to define Q_n and Ω_n. Let $\dot{\psi}_{ig}(\beta) = \frac{\partial}{\partial \beta} \mathbb{E}\, \psi(X_{ig}, \beta)$ and $\tilde{\psi}_g(X_g, \beta) := \sum_{i=1}^{n_g} \psi_{ig}(X_{ig}, \beta)$. Then, we define

$$Q_n := \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \dot{\psi}_{ig}(\beta) \tag{7}$$

$$\Omega_n := \frac{1}{n} \sum_{g=1}^{G} \mathbb{E}\, \tilde{\psi}_g(X_g, \beta) \tilde{\psi}_g(X_g, \beta)' \tag{8}$$

Assumption 2. *We assume that there exists a unique $\beta \in \mathcal{B}$ such that $\mathbb{E}\, \psi(X_{ig}, \beta) = 0$. We further assume that $\mathbb{E}\, \psi(X_{ig}, b)$ is continuously differentiable in b with continuous and invertible derivative matrix $\dot{\psi}_{ig}(b)$, for all b in an open neighbourhood N_β which contains β, and $\sup_{i,g} \mathbb{E}\, \|\psi(X_{ig}, \beta)\|^2 < \infty$. The estimator $\hat{\beta}$ satisfies*

$$\frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \psi(X_{ig}, \hat{\beta}) = o_P(\|\Omega_n\|^{1/2}) \tag{9}$$

and we assume that Ω_n and Q_n are full-rank, with the minimum eigenvalue $\lambda_{min}(\Omega_n)$ greater than or equal to $c > 0$ for all n. Lastly, we assume

$$(Q_n' \Omega_n^{-1} Q_n)^{1/2} \sqrt{n}(\hat{\beta} - \beta) = (Q_n' \Omega_n^{-1} Q_n)^{1/2} \frac{1}{\sqrt{n}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (Q_n')^{-1} \psi(X_{ig}, \beta) + o_P(1) \Rightarrow \mathcal{N}(0, I_q) \tag{10}$$

These assumptions are commonly satisfied in applications of interest in economics. The assumptions as stated serve the purpose of clarifying the environment: inference settings where asymptotically normal inference is correct in large samples.

Example 1 (*Efficient Smooth-GMM*)**.** If $\psi(X_i, \beta)$ is twice-continuously differentiable in β, $\mathbb{E}\, \frac{\partial}{\partial \beta} \psi(X_i, \beta)$ has full-rank, $\mathbb{E}\, \|\psi(X_i, \beta)\|^2 < \infty$, and the data are i.i.d., the efficient-GMM estimator $\hat{\beta}$ will generally satisfy (10) with

$$Q_n = \mathbb{E}\, \frac{\partial}{\partial \beta} \psi(X_i, \beta) = \dot{\psi}(\beta) \tag{11}$$

$$\Omega_n = \mathbb{E}\, \psi(X_i, \beta) \psi(X_i, \beta)' = \Psi(\beta) \tag{12}$$

See Newey and McFadden (1994), Theorem 3.2 for a complete list of the technical conditions.

Example 2 (*Cluster-Robust Inference*)**.** Consider the standard linear regression model:

$$y_{ig} = x_{ig}' \beta + \varepsilon_{ig} \tag{13}$$

where now observation i is in cluster g. Here, $\psi(X_{ig}, \beta) = x_{ig}(y_{ig} - x_{ig}'\beta) = x_{ig}\varepsilon_{ig}$. Suppose that scores are uncorrelated across clusters, in that $\mathbb{E}\, x_{ig} x_{jh}' \varepsilon_{ig} \varepsilon_{jh} = 0$ for $g \neq h$, but possibly within cluster $\mathbb{E}\, x_{ig} x_{jg}' \varepsilon_{ig} \varepsilon_{jg} \neq$

0. For the r in (6), we assume that there exists $s > r$, such that $\sup_{i,g} \mathbb{E}\,\|x_{ig}\|^{2s}, \sup_{i,g} \mathbb{E}\, y_{ig}^{2s} < \infty$. In this case, we have

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\, x_{ig} x_{ig}' \tag{14}$$

$$\Omega_n = \frac{1}{n}\sum_{g=1}^{G} \mathbb{E}\left[\left(\sum_{i=1}^{n_g} x_{ig}\varepsilon_{ig}\right)\left(\sum_{i=1}^{n_g} x_{ig}\varepsilon_{ig}\right)'\right]. \tag{15}$$

Then, if $\hat{\beta}$ is the OLS estimator of β and Q_n and Ω_n are full rank, then (10) is satisfied.

Example 3 (*Quantile Regression*)**.** Consider the quantile regression model:

$$y_i = x_i'\beta(\tau) + \varepsilon_i(\tau). \tag{16}$$

In this setting, $\psi(X_i, \beta(\tau)) = x_i\left(\mathbb{1}_{[y_i - x_i'\beta(\tau) < 0]} - \tau\right)$, and

$$\dot{\psi}_i(\beta(\tau)) = \mathbb{E}\, x_i x_i' f_\tau(\varepsilon_i(\tau)|x_i) \tag{17}$$

where $f_\tau(\cdot|x_i)$ is the conditional density of $\varepsilon_i(\tau)$ given x_i. Under i.i.d. sampling, we have that when $\sup_i \sup_u f_\tau(u|x_i) \leq C < \infty$, $\mathbb{E}\,\|x_i\|^2 < \infty$, and for

$$Q_n = \dot{\psi}_i(\beta(\tau)) \tag{18}$$

$$= \mathbb{E}\, x_i x_i' f_\tau(0|x_i) \tag{19}$$

$$\Omega_n = \mathbb{E}\, x_i x_i'\left(\mathbb{1}_{[y_i - x_i'\beta(\tau) < 0]} - \tau\right)^2, \tag{20}$$

then (10) is satisfied.

We will consider Wald statistics formed using plug-in estimators of Q_n and Ω_n:

$$\hat{Q}_n = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \hat{\dot{\psi}}_{ig}(\hat{\beta}) \tag{21}$$

$$\hat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G} \hat{\Psi}_g(\hat{\beta}) \tag{22}$$

Note that in some cases the functions $\dot{\psi}_{ig}$ will have to be estimated as well, as is the case in our quantile regression example. We also allow for a different choice other than $\tilde{\psi}_g$ in estimating Ω_n, which allows for applications where variants on the plug-in variance estimator are used. Our main requirement in both cases

is that:

$$\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\beta) - Q_n = o_P(1) \tag{23}$$

$$\frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\beta) - \Omega_n = o_P(1) \tag{24}$$

and thus we will be in a setting where W_n will lead to asymptotically valid inference. We will require a stronger assumption here, which implies smoothness on the estimation of Q_n and Ω_n in β. In a slight abuse of notation, in the following assumption we write $\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta})$ in place of $\mathbb{E}\,\hat{\dot{\psi}}_{ig}(b)\mid_{b=\hat{\beta}}$, and similarly for $\mathbb{E}\,\hat{\Psi}_g(\hat{\beta})$.

Assumption 3. *We assume that for any r_{1n} and r_{2n} such that*

$$r_{1n}\left(\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta)\right) = O_P(1),\ r_{1n}(\hat{\beta}-\beta) = O_P(1) \tag{25}$$

$$r_{2n}\left(\frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta)\right) = O_P(1),\ r_{2n}(\hat{\beta}-\beta) = O_P(1) \tag{26}$$

we have that

$$r_{1n}\left(\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\hat{\beta}) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta}) - \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta)\right) = o_P(1) \tag{27}$$

$$r_{2n}\left(\frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\hat{\beta}) - \mathbb{E}\,\hat{\Psi}_g(\hat{\beta}) - \frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta)\right) = o_P(1). \tag{28}$$

In characterising the asymptotic distributions of $\hat{Q}_n$ and $\hat{\Omega}_n$, we cannot rule out that the estimation of β does not impact the asymptotic behaviour of these matrices, which take $\hat{\beta}$ as an input. This assumption will imply that the estimation of β only enters in the first-order asymptotic behaviour of $\hat{Q}_n$ and $\hat{\Omega}_n$ through $\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta})$ and $\mathbb{E}\,\hat{\Psi}_g(\hat{\beta})$, which can effectively be handled using the delta-method. Furthermore, the additional noise due to estimating must go to zero at a faster rate than the estimation error in estimating Q_n or Ω_n if β was known. This condition is typically satisfied when the $\hat{\dot{\psi}}$ and $\hat{\Psi}$ functions are sufficiently smooth to satisfy a stochastic-equicontinuity condition. Differentiable functions, members of Donsker classes, and Lipschitz functions will generally satisfy these conditions. In the case of the linear models (Example 2), these requirements are almost trivially satisfied. The cases of GMM and quantile regression deserves a bit more attention.

Example 1, continued (*Efficient Smooth-GMM*)**.** In the case of smooth GMM settings when $\psi(X_i,\beta)$ is sufficiently smooth, both $\mathbb{E}\,\frac{\partial}{\partial\beta}\psi(X_i,\beta) = \dot{\psi}(\beta)$ and $\mathbb{E}\,\psi(X_i,\beta)\psi(X_i,\beta)' = \Psi(\beta)$ are differentiable, and Assumption 3 will be satisfied.

Assumption 3^{gmm} (*Smoothness assumption for GMM*)**.** *Let there exist a neighbourhood N_β of β such that*

$\psi(X_i, \beta)$ is twice-continuously differentiable with bounded derivatives in expectation

$$\mathbb{E}\left\|\frac{\partial}{\partial b}\psi(X_i, b)\right\|^2, \sum_{k=1}^{p}\mathbb{E}\left\|\frac{\partial^2}{\partial b \partial b'}\psi_k(X_i, b)\right\|^2 < \infty \tag{29}$$

for all $b \in N_\beta$, where $N_\beta \subset K_\beta$, K_β-compact.

Under this condition, Assumption 3 will be satisfied.

Lemma 1. *Under Assumption 3^{gmm}, we have that*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}\psi(X_i, \hat{\beta}) - \dot{\psi}(\hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}\psi(X_i, \beta) - \dot{\psi}(\beta) = o_P(1/\sqrt{n}) \tag{30}$$

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i, \hat{\beta})\psi(X_i, \hat{\beta})' - \Psi(\hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n}\psi(X_i, \beta)\psi(X_i, \beta)' - \Psi(\beta) = o_P(1/\sqrt{n}) \tag{31}$$

Example 3, continued (*Quantile Regression*)**.** In the case of quantile regression, we have to estimate:

$$Q_n = \mathbb{E}\, x_i x_i' f(0|x_i) \tag{32}$$

A standard estimator is a kernel estimator. Let $K(u)$ be a symmetric function of bounded variation such that $\int_{\mathbb{R}} uK(u)du = 0$ and $\int_{\mathbb{R}} u^2K(u)du = 1$. A kernel estimator of Q_n is given by

$$\hat{Q}_n := \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \frac{1}{h_n} K\left(\frac{y_i - x_i'\hat{\beta}_\tau}{h_n}\right) \tag{33}$$

where h_n is a bandwidth. Thus, we have that, in the notation we have used thus far, $\hat{\Psi}_i(\hat{\beta}_\tau) = x_i x_i' h_n^{-1} K((y - x_i'\hat{\beta}_\tau)/h_n)$. Assumption 3 is then satisfied with the following assumptions on K and the density of the error $\varepsilon_i(\tau)$:

Assumption 3^{qreg} (*Conditional density estimation for quantile regression*)**.**

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

$$\int_{\mathbb{R}} uK(u)du = 0, \qquad \int_{\mathbb{R}} u^2K(u)du = 1 \tag{34}$$

- *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 $\mathbb{E}(G_2(x_i)\|x_i\|^2) < \infty$.*
- $h_n = o(\log n/\sqrt{n})$.

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

The assumption on the integrability of envelope functions G_j for the density and its derivatives 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 effective envelope functions that are used in the bounds.

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), $n^2 h_n \to \infty$. In the estimation of $\mathbb{E}\, x_i x_i' f(0|x_i)$, we can decompose our kernel estimator into three components:

$$\hat{Q}_n = \frac{1}{n}\sum_{i=1}^n x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) \tag{35}$$

$$= \frac{1}{n}\sum_{i=1}^n x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) \tag{36}$$

$$+ \frac{1}{n}\sum_{i=1}^n x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) \tag{37}$$

$$- \left(\frac{1}{n}\sum_{i=1}^n x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right)\right) \tag{38}$$

$$+ \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right). \tag{39}$$

Standard results on asymptotic normality of kernel density estimators give that, for any constant matrix A,

$$\sqrt{nh_n}\left(\frac{1}{n}\sum_{i=1}^n \operatorname{tr}\left(A\left(x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right)\right)\right)\right) \Rightarrow \mathcal{N}(0, V(A)) \tag{40}$$

for $V(A) = \mathbb{E}[(x_i' A x_i)^2 f(0|x_i) R_K]$ and $R_K = \int K(u)^2 du$ is the roughness of the kernel $K(\cdot)$. Assumption 3^{qreg} implies that this is the only component of the asymptotic distribution of $\hat{Q}_n$ which is relevant for the first-order theory:

Lemma 2. *Under Assumption 3^{qreg}, we have that*

$$\frac{1}{n}\sum_{i=1}^n x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) \tag{41}$$

$$- \left(\frac{1}{n}\sum_{i=1}^n x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right)\right) = o_P(1/\sqrt{nh_n}) \tag{42}$$

$$\mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) = o_P(1/\sqrt{nh_n}) \tag{43}$$

Thus, the quantile regression model satisfies Assumption 3, where we can observe that the important rates to consider are $r_{1n} = O(\sqrt{nh_n})$ and $r_{2n} = O(\sqrt{n})$.

We are now ready to state our general result on the first-order asymptotics of W_n under fixed alternatives. First, so that the statement of theorem is focused on the key aspects, we define a sequence of constants B_n

and Ξ_n. Define:

$$\tilde{P}_n := (L'V_nL)^{-1}\Delta\Delta'(L'V_nL)^{-1} \tag{44}$$

$$P_{1n} := -L(2\tilde{P}_n - \text{diag}(\tilde{P}_n))L' \tag{45}$$

$$P_{2n} := (Q_n')^{-1}P_{1n}Q_n^{-1} \tag{46}$$

$$P_{3n} := -2V_nP_{1n}Q_n^{-1} \tag{47}$$

$$\tilde{p}_{1n} := -2\Delta'(L'V_nL)^{-1}L' \tag{48}$$

$$\tilde{p}_{2n} := \frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{p}\frac{\partial}{\partial\beta}\,\mathbb{E}\,\hat{\Psi}_{gk}(\beta)'[P_{2n}]_k \tag{49}$$

$$\tilde{p}_{3n} := \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p}\frac{\partial}{\partial\beta}\,\mathbb{E}\,\hat{\dot{\psi}}_{igk}(\beta)'[P_{3n}]_k \tag{50}$$

$$p_n := Q_n^{-1}(\tilde{p}_{1n} + \tilde{p}_{2n} + \tilde{p}_{3n}) \tag{51}$$

where $\hat{\Psi}_{gk}$ denotes the k^{th} row of $\hat{\Psi}_g$, $\hat{\dot{\psi}}_{igk}$ is the k^{th} row of $\hat{\dot{\psi}}_{ig}$, and $[P_{ln}]_k$ is the k^{th} column of P_{ln}. Then, we can define:

$$B_n := \frac{1}{n}\sum_{g=1}^{G}\text{tr}(P_{2n}(\Psi_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta))) + \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\text{tr}(P_{3n}(\dot{\psi}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta))) \tag{52}$$

$$Y_{gn} := \begin{pmatrix} \sum_{i=1}^{n_g} p_n'\psi(X_{ig},\beta) \\ \text{tr}(P_{2n}(\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta))) \\ \sum_{i=1}^{n_g}\text{tr}(P_{3n}(\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta))) \end{pmatrix} \tag{53}$$

$$\Xi_n := \frac{1}{n^2}\sum_{g=1}^{G}\mathbb{E}(\mathbf{1}'Y_{gn})^2 \tag{54}$$

Notice that $n\sqrt{\Xi_n} \to \infty$, as we require the number of clusters G to be increasing, and $\Xi_n = 0$ whenever $\Delta = 0$.

Theorem 1. *Assume that Assumptions 1-3 are satisfied. Suppose that for the $r \geq 2$ in Assumption 1 that there exists an $\epsilon > 0$ such that:*

$$\sup_{i,g}\mathbb{E}\,\|\psi(X_{ig},\beta)\|^{r+\epsilon}, \sup_{i,g}\mathbb{E}\left\|\hat{\dot{\psi}}_{ig}(\beta)\right\|^{r+\epsilon}, \sup_{g}\mathbb{E}\left\|\hat{\Psi}_g(\beta)\right\|^{r+\epsilon} < \infty \tag{55}$$

Then, for the previously defined sequences Ξ_n, B_n, where $\Xi_n > 0$, we have that

$$\frac{\frac{1}{n}W_n - \xi_n^2 - B_n}{\sqrt{\Xi_n}} \Rightarrow \mathcal{N}(0,1) \tag{56}$$

The additional moment conditions assumed here imply the uniform integrability conditions necessary for a central limit theorem to hold for the appropriate averages of $\psi(X_{ig},\beta)$, $\hat{\dot{\psi}}_{ig}(\beta)$, and $\hat{\Psi}_g(\beta)$. These conditions will typically be strictly stronger than those required for asymptotic normality of $\hat{\beta}$. When standard plug-in estimators for Q_n and Ω_n are used, then there will be no bias term B_n, however different choices of estimators of Q_n and Ω_n will lead to different centering sequences B_n, as well as different normalizations Ξ_n. This highlights an important feature of these asymptotics: under fixed alternatives, the choice of variance

estimator impacts the limit distribution of the test statistic.

We now discuss, and derive expressions for, B_n and Ξ_n in the particular cases of cluster robust inference, quantile regression, and IV.

Example 2, continued (*Cluster-Robust Inference*)**.** In this case we specialise to the case when $L = \ell \in \mathbb{R}^p$. The linear setting, with plug-in estimators, implies $B_n = 0$. We also can simplify many of the expressions making up the components of Ξ_n. Note that in this case:

$$\tilde{P}_n = (L'V_nL)^{-1}\Delta\Delta'(L'V_nL)^{-1} \tag{57}$$

$$= \frac{(\ell'(\beta - \beta_0))^2}{(\ell V_n \ell)^2} \tag{58}$$

$$= \frac{\xi_n^2}{\ell' V_n \ell} \tag{59}$$

$$P_{1n} = -\ell\ell' \frac{\xi_n^2}{\ell' V_n \ell} \tag{60}$$

This simplification implies that for:

$$b_n = \frac{\Delta(Q_n')^{-1}\ell}{\ell' V_n \ell} \tag{61}$$

$$c_n = \frac{2\Delta V_n \ell}{\ell' V_n \ell} \tag{62}$$

$$a_n = -\frac{2\Delta}{\ell' V_n \ell} Q_n^{-1}\ell - \frac{2}{n} Q_n^{-1} \sum_{g=1}^{G} \mathbb{E}\left[\left(\sum_{i=1}^{n_g} x_{ig}x_{ig}'\right) b_n b_n' \left(\sum_{i=1}^{n_g} x_{ig}\varepsilon_{ig}\right)\right] \tag{63}$$

In this case, Y_{gn} contains a term that is linear in the $\psi_{ig}(X_{ig}, \beta)$, a quadratic form in the $\Psi_g(\beta) = \mathbb{E}(\sum_{i=1}^{n_g} x_{ig}\varepsilon_{ig})(\sum_{i=1}^{n_g} x_{ig}\varepsilon_{ig})'$, and a bilinear form in the $\dot{\hat{\psi}}_{ig}(\beta) = \mathbb{E}\, x_{ig}x_{ig}'$:

$$Y_{gn} = \begin{pmatrix} \sum_{i=1}^{n_g} a_n' x_{ig}\varepsilon_{ig} \\ \left(\sum_{i=1}^{n_g} b_n' x_{ig}\varepsilon_{ig}\right)^2 - \mathbb{E}\left(\sum_{i=1}^{n_g} b_n' x_{ig}\varepsilon_{ig}\right)^2 \\ \sum_{i=1}^{n_g} b_n' x_{ig}x_{ig}' c_n - \sum_{i=1}^{n_g} \mathbb{E}(b_n x_{ig}x_{ig}' c_n) \end{pmatrix} \tag{64}$$

The result in Theorem 1 then specialises to:

$$\frac{\frac{(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{V}_n\ell} - \xi_n^2}{\sqrt{\Xi_n}} \Rightarrow \mathcal{N}(0,1) \tag{65}$$

Now suppose that there is no cluster dependence, i.e. $\Omega_n = \mathbb{E}\, x_{ig}x_{ig}'\varepsilon_{ig}^2$, however a cluster-robust variance estimator is still used. Inspection of Y_{gn} tells us that the main difference is in the second element of the vector

$$\tilde{Y}_{gn} = \begin{pmatrix} \sum_{i=1}^{n_g} a_n' x_{ig}\varepsilon_{ig} \\ \sum_{i=1}^{n_g} (b_n' x_{ig}\varepsilon_{ig})^2 - \sum_{i=1}^{n_g} \mathbb{E}(b_n' x_{ig}\varepsilon_{ig})^2 \\ \sum_{i=1}^{n_g} b_n' x_{ig}x_{ig}' c_n - \sum_{i=1}^{n_g} \mathbb{E}(b_n x_{ig}x_{ig}' c_n) \end{pmatrix} \tag{66}$$

which is Y_{gn} specialised to the case that $\hat{\Omega}_n$ reduces to the standard heteroskedastic-robust variance estimator when there is presumed no cluster dependence. Under this assumption that the observations are i.i.d., i.e.

no cluster-dependence, we have that under the null hypothesis, $\ell'\beta = \theta_0$,

$$\frac{n(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{Q}_n^{-1}\hat{\Omega}_n\hat{Q}_n^{-1}\ell} \Rightarrow \chi_1^2 \tag{67}$$

$$\frac{n(\ell'\hat{\beta} - \theta_0)^2}{\ell'\hat{Q}_n^{-1}\tilde{\Omega}_n\hat{Q}_n^{-1}\ell} \Rightarrow \chi_1^2. \tag{68}$$

Thus, both statistics provide for valid inference under the null hypothesis, and have the same first-order asymptotic behaviour in that case. Let $\tilde{\Xi}_n := \frac{1}{n^2}\sum_{g=1}^{G}\mathbb{E}(\mathbf{1}'\tilde{Y}_{gn})^2$. Comparing Ξ_n to $\tilde{\Xi}_n$ gives us an insight into differences in the asymptotic behaviour between the two procedures.

Proposition 1. *In the linear regression model, suppose a test of $H_0 : \ell'\beta = \theta_0$ can be conducted using both a cluster robust variance estimator $\hat{\Omega}_n$ and the standard heteroskedasticity-robust variance estimator $\tilde{\Omega}_n$:*

$$\hat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G}\left(\sum_{i=1}^{n_g} x_{ig}\hat{\varepsilon}_{ig}\right)\left(\sum_{i=1}^{n_g} x_{ig}\hat{\varepsilon}_{ig}\right)' \tag{69}$$

$$\tilde{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}x_{ig}'\hat{\varepsilon}_{ig}^2 \tag{70}$$

Then, under any fixed alternative $\theta \neq \theta_0$,

$$\frac{\frac{(\ell'\hat{\beta}-\theta_0)^2}{\ell'\hat{Q}_n^{-1}\hat{\Omega}_n\hat{Q}_n^{-1}\ell} - \xi_n^2}{\sqrt{\Xi}_n} \Rightarrow \mathcal{N}(0,1) \tag{71}$$

$$\frac{\frac{(\ell'\hat{\beta}-\theta_0)^2}{\ell'\hat{Q}_n^{-1}\tilde{\Omega}_n\hat{Q}_n^{-1}\ell} - \xi_n^2}{\sqrt{\tilde{\Xi}}_n} \Rightarrow \mathcal{N}(0,1) \tag{72}$$

with $\Xi_n \geq \tilde{\Xi}_n$, with equality if and only if $\mathrm{Var}(x_{ig}\varepsilon_{ig}) = 0$.

Example 3, continued (*Quantile Regression*)**.** In the case of quantile regression we also obtain significant simplification, this time because the Q_n matrix is estimated at a much slower rate than β and Ω_n. The reason for this is that Q_n is estimated at the $\sqrt{nh_n}$ rate, as is typical for kernel density estimators, while $\sqrt{n}(\hat{\beta} - \beta)$, $\sqrt{n}(\hat{\Omega}_n - \Omega_n) = O_P(1)$. We again focus on the case $L = \ell \in \mathbb{R}^p$, so that we are testing the hypothesis $H_0 : \ell'\beta = \theta_0$. This is also an example of when the bias term B_n is necessary to characterise the asymptotic theory. We have the same simplification as when we considered linear hypotheses in the linear model

$$b_n = \frac{(Q_n')^{-1}\ell}{\sqrt{\ell' V_n \ell}} \tag{73}$$

$$c_n = \frac{2V_n\ell}{\sqrt{\ell' V_n \ell}}. \tag{74}$$

Here, the bias term is:

$$B_n = b_n' \, \mathbb{E}\left(x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i}{h_n} \right) \right) c_n \tag{75}$$

$$= \frac{1}{2} b_n' \, \mathbb{E}(x_i x_i' f''(0|x_i) h_n^2) c_n \tag{76}$$

$$\tag{77}$$

Proposition 2. *Recall that R_K is the roughness of the kernel K, and $f(\cdot|x_i)$ and $f''(\cdot|x_i)$ are the conditional density of the error and the second derivative respectively. The, for the quantile regression model, we have that*

$$\frac{\sqrt{nh_n}\left(\frac{1}{n}W_n - \xi_n^2 - B_n\right)}{\sqrt{\mathbb{E}\left[(b_n' x_i x_i' c_n)^2 f(0|x_i) R_K\right]}} \Rightarrow \mathcal{N}(0,1). \tag{78}$$

A direct implication of this result is that the deviations of the test statistic from the non-centrality parameter in large samples are driven by the nonparametric estimator of the variance component Q_n. This is quite different from the local asymptotic theory which treats all estimators of Q_n as equivalent, as long as they are consistent. Here, the choice of kernel and bandwidth h_n play a role in the asymptotic behaviour of the test statistic.

4 Power Calculations and Simulations

We can use the central limit-theorem result in Theorem 1 to approximate power, just as we did in Section 2. Using (1) as a guide, consider the rejection probability of the standard Wald test: letting q_γ be the γ-quantile of the appropriate χ^2-distribution, we have that the power of a level-α test is given by:

$$P(W_n > q_{1-\alpha}) = P\left(\frac{\frac{1}{n}W_n - \xi_n^2}{\sqrt{\Xi_n}} > \frac{q_{1-\alpha}}{n\sqrt{\Xi_n}} - \frac{\xi_n^2}{\sqrt{\Xi_n}} \right) \tag{1}$$

Since $\frac{1}{n}W_n - \xi_n^2 \xrightarrow{P} 0$, and the lefthand side of (1) is a consistent test, we must have that $\xi_n^2/\sqrt{\Xi_n} \to \infty$. Thus, The only way we can approximate the power is by specifying a sequence of critical values that tends to infinity as well. Consider any sequence of the form:

$$C_n^* = n\left(\xi_n^2 - t_n \sqrt{\Xi_n} + \frac{C_n}{n} \right) \tag{2}$$

If $\Delta = \theta - \theta_0 = 0$, then note that $C_n^* = C_n$. Otherwise, rearranging the expression, we see that:

$$P\left(W_n > C_n^*\right) = P\left(\frac{\frac{1}{n}W_n - \xi_n^2}{\sqrt{\Xi_n}} > -t_n + \frac{C_n}{n\sqrt{\Xi_n}} \right) \tag{3}$$

Thus, we have that $P(W_n > C_n^*) - \Phi(t_n) \to 0$. We can interpret $C_n/n\sqrt{\Xi_n}$ as a kind of continuity-correction that allows us to include the case $\Delta = 0$ which would imply $\Xi_n = 0$, in which case $P(W_n > C_n)$ converges to an upper-tail probability of a chi-square random variable.

One way to use these approximations to calculate power is to set $C_n^* = C_n = q_{1-\alpha}$. This leads to choosing $t_n = \xi_n^2/\sqrt{\Xi_n}$. This will lead to the approximation $P(W_n > q_{1-\alpha}) \approx \Phi(\xi_n^2/\sqrt{\Xi_n} - q_{1-\alpha}/n\sqrt{\Xi_n})$

N	G	Size	Power	Relative Error
56	5	5.57E-01	0.711	-0.296
125	10	2.53E-01	0.631	-0.207
196	15	1.28E-01	0.599	-0.165
268	20	6.98E-02	0.578	-0.135
344	25	3.63E-02	0.567	-0.118
423	30	2.01E-02	0.571	-0.125
502	35	1.13E-02	0.561	-0.108
583	40	5.75E-03	0.552	-0.094
664	45	3.38E-03	0.549	-0.090
747	50	2.10E-03	0.545	-0.083
831	55	9.70E-04	0.548	-0.087
916	60	5.70E-04	0.548	-0.087
1001	65	3.70E-04	0.547	-0.085
1089	70	2.30E-04	0.535	-0.066
1175	75	9.00E-05	0.535	-0.066
1262	80	5.00E-05	0.542	-0.078

Table 1: Asymptotic Power = 0.50

when $\Delta \neq 0$. Note that this will only lead to predicted power over 1/2 asymptotically. This is akin to using local-asymptotic power analysis in finite samples, in that for this choice of t, the approximation is not asymptotically valid, but it is a translation of the asymptotic framework to the practical question of power analysis. In the case of local power, the approximation works well when the alternatives of interest are small.

Another option is to consider convergence along a sequence of critical values which is indeed diverging off to ∞. If we choose $C_n = n\sqrt{\Xi_n}$ and $t_n = 1 - q_\beta$, then we have that $P(W_n > C_n^*) \to 1 - \beta$. These asymptotics tell the following story: for a given alternative $\theta = \theta_0 + \Delta$, we can approximate power as size converges to zero. Thus, it allows practitioners to approximate a sequence of pairs (n, α) such that power is approximately β along the sequence.

4.1 Cluster-Robust Regression Simulations

For our simulations, we focus on the linear regression model:

$$y_{ig} = x_{ig}'\beta + u_g + \varepsilon_{ig} \tag{1}$$

where $\varepsilon_{ig} \sim \mathcal{N}(0, \sqrt{1-\rho}\sigma^2)$, $u_g \sim \mathcal{N}(0, \sqrt{\rho}\sigma^2)$, $x_{ig} = (1, z_{ig}')'$, with $\mathbb{E}\, z_{ig} z_{ig}' = I$. The design follows MacKinnon et al. (2023), including how we let clusters grow with the number of clusters. We set asymptotic power to be 0.5, 0.80, and 0.90. We test the hypothesis $\beta_1 = 0$ against a two-sided alternative. For each sample size, we computed the size of a test using the critical value C_n^* when $\beta_1 = 0$, the power, and then the relative error of the power estimates.

The simulation results are summarised in the Tables 1-3. We removed combinations of sample size and asymptotic power which lead to size or power in simulations equal to 1.

In general, the approximation improves for a given sample size as we increase the target power. For lower values of power, we see that the size of the test and sample size get quite small without the relative error in the approximation getting below 6%, however for power at 0.90, size is only 0.0269 for the approximation to be accurate in this range. This implies that power calculations using our approach will generally be useful in large samples, for power in the range generally useful for power analysis and sample size calculations: at

N	G	Size	Power	Relative Error
196	15	7.20E-01	0.944	-0.152
268	20	4.27E-01	0.906	-0.117
344	25	2.64E-01	0.891	-0.102
423	30	1.68E-01	0.884	-0.095
502	35	1.05E-01	0.876	-0.087
583	40	6.66E-02	0.875	-0.085
664	45	4.27E-02	0.869	-0.079
747	50	2.72E-02	0.866	-0.076
831	55	1.75E-02	0.856	-0.065
916	60	1.07E-02	0.858	-0.067
1001	65	6.80E-03	0.856	-0.065
1089	70	4.63E-03	0.852	-0.061
1175	75	2.73E-03	0.851	-0.059
1262	80	1.61E-03	0.851	-0.059

Table 2: Asymptotic Power = 0.80

N	G	Size	Power	Relative Error
502	35	5.29E-01	0.986	-0.087
583	40	3.27E-01	0.978	-0.080
664	45	2.13E-01	0.975	-0.077
747	50	1.39E-01	0.972	-0.074
831	55	9.03E-02	0.963	-0.066
916	60	6.12E-02	0.963	-0.066
1001	65	3.96E-02	0.959	-0.062
1089	70	2.69E-02	0.956	-0.059
1175	75	1.74E-02	0.959	-0.062
1262	80	1.23E-02	0.957	-0.059

Table 3: Asymptotic Power = 0.90

least 0.80 for size in a typical range of 0.001-0.05.

5 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. This opens up the possibility of comparing different variance estimators with testing in mind. This contrasts with previous comparisons that have previously been made 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 justification for assuming there is a cost for clustering at too-coarse a level.

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 power analysis could provide guidance for tuning parameter choices in that setting. We could also extend the results to accomodate settings where the variance estimator is not asymptotically normal, as is the case when the 4^{th} moment of

the influence function does not exist. The theory in Bentkus et al. (2007) introduces similar results in the case of the 1-sample t-test, and similar results could be produced in a more general setting, such as in the case of linear processes driven by i.i.d. infinite-variance innovations (see, for example, Cavaliere, Georgiev, and Taylor (2016)).

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica 88*(1), 265–296.

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

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics 38*(2), 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*(2), 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*(1), 249–275.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Canay, I. A. and T. Otsu (2012). Hodges–Lehmann optimality for testing moment conditions. *Journal of Econometrics 171*(1), 45–53.

Cavaliere, G., I. Georgiev, and A. R. Taylor (2016). Sieve-based inference for infinite-variance linear processes.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014, August). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics 42*(4), 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), 1–37.

Engle, R. F. (1984). Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, Volume 2, pp. 775–826. Elsevier.

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

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics 210*(2), 268–290.

Hodges, J. and E. Lehmann (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics 27*(2), 324–335.

Iacone, F., S. J. Leybourne, and A. R. Taylor (2013). On the behavior of fixed-b trend break tests under fractional integration. *Econometric Theory 29*(2), 393–418.

Kato, K. (2012). Asymptotic normality of Powell's kernel estimator. *Annals of the Institute of Statistical Mathematics 64*(2), 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*(1), 26–51.

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

Le Cam, L. (2012). *Asymptotic Methods in Statistical Decision Theory*. Springer Science & Business Media.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023). Testing for the appropriate level of clustering in linear regression models. *Journal of Econometrics*.

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*(100), 41–53.

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.

Shao, Q. and R. Zhang (2009). Asymptotic distributions of non-central studentized statistics. *Science in China, Series A: Mathematics 52*(6), 1262–1284.

Sun, Y., P. C. Phillips, and S. Jin (2008). Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica 76*(1), 175–194.

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

A Proofs of Main Results

A.1 Proof of Theorem 1

Proof. We start by deriving the asymptotic expansion (4).

$$(L'\hat{\beta} - \theta_0)'(L'\hat{V}_n L)^{-1}(L'\hat{\beta} - \theta_0) - \xi_n^2 = (L'\hat{\beta} - \theta)'(L'\hat{V}_n L)^{-1}(L'\hat{\beta} - \theta) \tag{1}$$

$$- 2\Delta'(L'\hat{V}_n L)^{-1}(L'\hat{\beta} - \theta) \tag{2}$$

$$+ \Delta'(L'\hat{V}_n L)^{-1}\Delta - \xi_n^2 \tag{3}$$

$$= L_{1n} + L_{2n} + L_{3n}. \tag{4}$$

$L_{1n} = O_P(1/n)$, and thus we will see that we can ignore it. L_{2n} is one of three terms which depends on $L'\hat{\beta} - \theta$, the other two which will come out of L_{3n}, which we will focus on now. First, we will find an expression which is a linear functional of $\hat{V}_n$:

$$\tilde{P}_{1n} := (L'V_n L)^{-1}\Delta\Delta'(L'V_n L)^{-1} \tag{5}$$

$$P_{1n} := -L(2\tilde{P}_{1n} - \text{diag}(\tilde{P}_{1n}))L' \tag{6}$$

$$\Delta'(L'\hat{V}_n L)^{-1}\Delta - \xi_n^2 = \text{tr}(P_{1n}(\hat{V}_n - V_n)) + R_{1n} \tag{7}$$

$$R_{1n} = o_P(\|L\|^2\|\hat{V}_n - V_n\|) = o_P(\|\hat{V}_n - V_n\|) \tag{8}$$

Now, recalling that $V_n = (Q_n'\Omega_n^{-1}Q_n)^{-1} = Q_n^{-1}\Omega_n(Q_n')^{-1}$, we can write:

$$\text{tr}(P_{1n}(\hat{V}_n - V_n)) = \text{tr}(P_{1n}(\hat{Q}_n^{-1}\hat{\Omega}_n(\hat{Q}_n')^{-1} - Q_n^{-1}\Omega_n(Q_n')^{-1})) \tag{9}$$

$$= \text{tr}((Q_n')^{-1}P_{1n}Q_n^{-1}(\hat{\Omega}_n - \Omega_n)) \tag{10}$$

$$- \text{tr}(2V_nP_{1n}V_nQ_n'\Omega_n^{-1}(\hat{Q}_n - Q_n)) \tag{11}$$

$$+ R_{2n} + R_{3n} \tag{12}$$

$$= \text{tr}(P_{2n}(\hat{\Omega}_n - \Omega_n)) + \text{tr}(P_{3n}(\hat{Q}_n - Q_n)) + R_{2n} + R_{3n} \tag{13}$$

$$R_{2n} = o_P(\|\hat{\Omega}_n - \Omega_n\|) \tag{14}$$

$$R_{3n} = o_P(\|\hat{Q}_n - Q_n\|) \tag{15}$$

Note that P_{3n} simplifies to $2V_nP_{1n}Q_n^{-1}$ Now, the estimators are of the form:

$$\hat{Q}_n = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\hat{\beta}) \tag{16}$$

$$\hat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\hat{\beta}) \tag{17}$$

Thus, in general, the asymptotic distribution of the estimators will depend on the asymptotic distribution of $\hat{\beta}$. We will first consider $\hat{Q}$. We decompose $\hat{Q}_n$ into three parts:

$$\hat{Q}_n = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\beta) \tag{18}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\hat{\beta}) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta}) - \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta) \tag{19}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta}) - \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta) \tag{20}$$

$$= J_{1n} + J_{2n} + J_{3n} \tag{21}$$

where with a (somewhat standard) abuse of notation, $\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta})$ is the expectation evaluated at $\hat{\beta}$, i.e. $\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\hat{\beta}) = \mathbb{E}\,\hat{\dot{\psi}}_{ig}(b)\mid_{b=\hat{\beta}}$. J_{1n} when properly centred by $\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta)$ and rescaled will be asymptotically normal following a Lindeberg CLT. J_{3n} will be asymptotically linear by the delta-method, and J_{2n} is negligible by assumption. The delta-method gives the additional term depending on $\hat{\beta}$: for any matrix A,

$$\operatorname{tr}(AJ_{2n}) = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p} a_k'\left[\frac{\partial\,\mathbb{E}\,\hat{\dot{\psi}}_k}{\partial\beta}(\beta)\right]_{ig}(\hat{\beta}-\beta) + o_P(\|\hat{\beta}-\beta\|) \tag{22}$$

where $\hat{\dot{\psi}}_j$ is the j^{th} row of $\hat{\dot{\psi}}$ and a_j is the j^{th} column of A. We obtain a similar expression for $\hat{\Omega}_n$:

$$\hat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\beta) \tag{23}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\hat{\beta}) - \mathbb{E}\,\hat{\Psi}_g(\hat{\beta}) - \frac{1}{n}\sum_{g=1}^{G}\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta) \tag{24}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\mathbb{E}\,\hat{\Psi}_g(\hat{\beta}) - \frac{1}{n}\sum_{g=1}^{G}\mathbb{E}\,\hat{\Psi}_g(\beta) \tag{25}$$

$$= H_{1n} + H_{2n} + H_{3n}. \tag{26}$$

In a similar fashion, for any matrix A, we have that:

$$\operatorname{tr}(AH_{2n}) = \frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{p} a_k'\left[\frac{\partial\,\mathbb{E}\,\hat{\Psi}_k}{\partial\beta}(\beta)\right]_{g}(\hat{\beta}-\beta) + o_P(\|\hat{\beta}-\beta\|) \tag{27}$$

where $\hat{\Psi}_k$ denotes the k^{th} row of $\hat{\Psi}$. Note that in light of these expansions, in this setting neither Q_n nor Ω_n can be estimated at a faster rate than β. This implies that $\hat{\beta}$ cannot dominate the (first-order) behaviour of the test statistic W_n under a fixed alternative. We now characterise the joint behaviour of $\hat{\beta}$, $\hat{Q}_n$, and $\hat{\Omega}_n$. Let $\tilde{Q}_n := \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbb{E}\,\hat{\dot{\psi}}(\beta)$ and $\tilde{\Omega}_n := \frac{1}{n}\sum_{g=1}^{G}\mathbb{E}\,\hat{\Psi}_g(\beta)$. For a sequence of constant matrices A_{1n} and

A_{2n}, define:

$$a_{1n} := \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p}[A_{1n}]_k' \left[\frac{\partial \mathbb{E}\,\hat{\dot{\psi}}_k}{\partial\beta}(\beta)\right]_{ig} \tag{28}$$

$$a_{2n} := \frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{p}[A_{2n}]_k' \left[\frac{\partial \mathbb{E}\,\hat{\Psi}_k}{\partial\beta}(\beta)\right]_{g} \tag{29}$$

where $[A_{jn}]_k$ denotes the k^{th} column of matrix A_{jn}. Let $\tilde{a}_n := (a_n + a_{1n} + a_{2n})$. Then, we have that:

$$a_n'(\hat{\beta}-\beta) + \mathrm{tr}(A_{1n}(\hat{Q}_n - \tilde{Q}_n)) + \mathrm{tr}(A_{2n}(\hat{\Omega}_n - \tilde{\Omega}_n)) \tag{30}$$

$$= \tilde{a}_n'(\hat{\beta}-\beta) \tag{31}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathrm{tr}(A_{1n}(\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta)) \tag{32}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\mathrm{tr}(A_{2n}(\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta)) + o_P(\|\hat{\beta}-\beta\|) \tag{33}$$

$$= \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\tilde{a}_n'(Q_n')^{-1}\psi(X_{ig},\beta) \tag{34}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathrm{tr}(A_{1n}(\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta)) \tag{35}$$

$$+ \frac{1}{n}\sum_{g=1}^{G}\mathrm{tr}(A_{2n}(\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta)) + o_P(\|\hat{\beta}-\beta\|) \tag{36}$$

If we then let

$$Y_{gn} := \begin{pmatrix} \sum_{i=1}^{n_g}\tilde{a}_n'(Q_n')^{-1}\psi(X_{ig}) \\ \mathrm{tr}(A_{2n}(\hat{\Psi}_g(\beta) - \mathbb{E}\,\hat{\Psi}_g(\beta))) \\ \sum_{i=1}^{n_g}\mathrm{tr}(A_{1n}(\hat{\dot{\psi}}_{ig}(\beta) - \mathbb{E}\,\hat{\dot{\psi}}_{ig}(\beta))) \end{pmatrix} \tag{37}$$

Then, if $\Xi_n := \frac{1}{n^2}\sum_{g=1}^{G}\mathbf{1}'\,\mathbb{E}\,Y_{gn}Y_{ng}'\mathbf{1}$, we have that, given the assumptions on moment existence,

$$\frac{a_n'(\hat{\beta}-\beta) + \mathrm{tr}(A_{1n}(\hat{Q}_n - \tilde{Q}_n)) + \mathrm{tr}(A_{2n}(\hat{\Omega}_n - \tilde{\Omega}_n))}{\sqrt{\Xi_n}} \Rightarrow \mathcal{N}(0,1) \tag{38}$$

from Theorem 2 of Hansen and Lee (2019). Plugging in the p_n and P_{jn} lead to the required result, with the bias terms given by:

$$B_n = \mathrm{tr}(P_{2n}(\tilde{\Omega}_n - \Omega_n) + \mathrm{tr}(P_{3n}(\tilde{Q}_n - Q_n)) \tag{39}$$

□

A.2 Proof of Proposition 1

Proof. Let $\Xi_0 = \mathbb{E}(\mathbf{1}'\tilde{Y}_{ng})^2$. When computing $\Xi_n - \Xi_0$, we obtain a great simplification that when looking at $\mathbb{E}\,Y_{gn}Y_{gn}' - \mathbb{E}\,\tilde{Y}_{gn}\tilde{Y}_{gn}'$, we note that all terms are zero, except for the second diagonal element. We can

compute this component as a function of the moments of

$$[\mathbb{E}\, Y_{gn}Y_{gn}']_{2,2} - [\mathbb{E}\, \tilde{Y}_{gn}\tilde{Y}_{gn}']_{2,2} = \sum_{i\neq j} \mathbb{E}(b_n'x_{ig}'\varepsilon_{ig})^2\, \mathbb{E}(b_n'x_{jg}\varepsilon_{jg})^2 \geq 0 \tag{1}$$

where we use the notation $[\cdot]_{a,b}$ to denote the (a,b) element of the relevant matrix. To see why (1) is true, notice that the components of $\tilde{Y}_{ng}$ and Y_{ng} are mean-zero. However, the variances are:

$$\operatorname{Var}\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right) = n_g \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig}) \tag{2}$$

When the n_g are asymptotically negligible, as is the case when Assumption 6 is true, then using a cluster-robust variance estimator poses no issue for validity of inference: the variance estimator is still consistent. However, now we must assess the variances of these variance estimators:

$$\operatorname{Var}\left(\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right)^2\right) = \mathbb{E}\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right)^4 - \left(\mathbb{E}\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right)^2\right)^2 \tag{3}$$

$$= \mathbb{E}\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right)^4 - n_g^2 \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{4}$$

We can write this fourth-moment in terms of the fourth cumulant, which we will denote $k_4(\cdot)$, and use the property that the cumulant of the sum is the sum of cumulants:

$$\mathbb{E}\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right)^4 = k_4\left(\sum_{i=1}^{n_g} b_n'x_{ig}\varepsilon_{ig}\right) + 3n_g^2 \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{5}$$

$$= \sum_{i=1}^{n_g} k_4(b_n'x_{ig}\varepsilon_{ig}) + 3n_g^2 \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{6}$$

$$= n_g k_4(b_n'x_{ig}\varepsilon_{ig}) + 3n_g^2 \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{7}$$

Thus, note that:

$$[\mathbb{E}\, Y_{gn}Y_{gn}']_{2,2} - [\mathbb{E}\, \tilde{Y}_{gn}\tilde{Y}_{gn}']_{2,2} = \operatorname{Var}\left(\left(\sum_{i=1}^{n_g} (b_n'x_{ig}\varepsilon_{ig})\right)^2\right) - \operatorname{Var}\left(\sum_{i=1}^{n_g} (b_n'x_{ig}\varepsilon_{ig})^2\right) \tag{8}$$

$$= n_g k_4(b_n'x_{ig}\varepsilon_{ig}) + 2n_g^2 \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{9}$$

$$\quad - n_g k_4(b_n'x_{ig}\varepsilon_{ig}) - 2n_g \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{10}$$

$$= 2n_g(n_g-1) \operatorname{Var}(b_n'x_{ig}\varepsilon_{ig})^2 \tag{11}$$

$$= \sum_{i\neq j} \mathbb{E}(b_n'x_{ig}\varepsilon_{ig})^2\, \mathbb{E}(b_n'x_{jg}\varepsilon_{jg})^2 \tag{12}$$

□

A.3 Proof of Proposition 2

Proof. As a result of Lemma 2, and since we know that $\hat{\beta} - \beta = O_P(1/\sqrt{n})$ and $\hat{\Omega}_n - \Omega_n = O_P(1/\sqrt{n})$. Standard results on kernel density estimation lead to the expression for B_n and that for any matrix A we have that:

$$\sqrt{nh_n}\left(\frac{1}{n}\sum_{i=1}^{n} b_n' x_i x_i' c_n \frac{1}{h_n} K\left(\frac{\varepsilon_i}{h_n}\right) - \xi^2 - B_n\right) \Rightarrow \mathcal{N}\left(0,\, \mathbb{E}(b_n' x_i x_i' c_n)^2 f(0|x_i) R_K\right) \tag{1}$$

□

B Proofs of Lemmas

B.1 Proof of Lemma 1

A Taylor expansion gives us that, for $\hat{\beta} \in N_\beta$.

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\beta}\psi(X_i,\hat{\beta}) - \dot{\psi}(\hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\beta}\psi(X_i,\beta) - \dot{\psi}(\beta) \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\frac{\partial^2}{\partial\beta\partial\beta'}\psi_k(X_i,\bar{\beta}) - \mathbb{E}\,\frac{\partial^2}{\partial\beta\partial\beta'}\psi_k(X_i,\bar{\beta})\right)'(\hat{\beta}-\beta) \tag{2}$$

$$= o_P(1)O_P(1/\sqrt{n}) = o_P(1/\sqrt{n}) \tag{3}$$

by the assumption of bounded derivatives and the fact that $\sqrt{n}(\hat{\beta}-\beta) = O_P(1)$. Since $P(\hat{\beta} \in N_\beta) \to 1$, for any neighborhood N_β, the result follows. Similarly, for

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i,\hat{\beta})\psi(X_i,\hat{\beta})' - \Psi(\hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n} \psi(X_i,\beta)\psi(X_i,\beta)' - \Psi(\beta) \tag{4}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} \psi(X_i,\bar{\beta})'\frac{\partial}{\partial\beta}\psi(X_i,\bar{\beta})' - \mathbb{E}\left(\psi(X_i,\bar{\beta})'\frac{\partial}{\partial\beta}\psi(X_i,\bar{\beta})'\right)\right)(\hat{\beta}-\beta) \tag{5}$$

$$= o_P(1/\sqrt{n}) \tag{6}$$

B.2 Proof of Lemma 2

Proof. Boundedness of the kernel function K and the existence of integrable functions G_j which uniformly bound the error density and the derivatives implies that

$$\mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) = O_P(1/\sqrt{n}) \tag{1}$$

via the delta-method. Now, we focus on:

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\hat{\varepsilon}_i(\tau)}{h_n}\right) \tag{2}$$

$$- \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right) - \mathbb{E}\, x_i x_i' \frac{1}{h_n} K\left(\frac{\varepsilon_i(\tau)}{h_n}\right)\right) \tag{3}$$

$$= \frac{1}{\sqrt{n}}(\mathbb{G}_n(xx'\hat{f}(\hat{\beta}) - \mathbb{G}_n(xx'\hat{f}(\beta))) \tag{4}$$

using the common notational shorthand $\mathbb{G}_n\psi(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,\beta) - \mathbb{E}\,\psi(X_i,\beta)$. An implication of Assumption 3^{qreg} 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,

$$\left|K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right| \leq \int \left|\mathbb{1}_{[\frac{\varepsilon_i - x_i'\delta_s}{h_n}, \frac{\varepsilon_i - x_i'\delta_t}{h_n}]}(x)\right| d(K_1(x) + K_2(x)) \tag{5}$$

Thus, using (5), we can use Hölder's inequality to bound the mean-squared difference:

$$\mathbb{E}\left[\left(K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right)^2 \middle| x_i\right] \tag{6}$$

$$\leq \int \mathbb{E}\left|\mathbb{1}_{[\frac{\varepsilon_i - x_i'\delta_s}{h_n}, \frac{\varepsilon_i - x_i'\delta_t}{h_n}]}(x)\right| d(K_1(x) + K_2(x))|K|_v \tag{7}$$

$$= \int \left|\int_{x_i'\delta_s + h_n x}^{x_i'\delta_t + h_n x} f(\varepsilon|x_i)d\varepsilon\right| d(K_1(x) + K_2(x))|K|_v \tag{8}$$

$$\leq \|f(\cdot|x_i)\|_\infty |K|_v^2 \|x_i\|_2 \|t - s\|_2 \tag{9}$$

Now, by Assumption 3^{qreg}:

$$\mathbb{E}\left[\left(K\left(\frac{\varepsilon_i - x_i'\delta_t}{h_n}\right) - K\left(\frac{\varepsilon_i - x_i'\delta_s}{h_n}\right)\right)^2\right] = O(\|t - s\|_2) \tag{10}$$

Putting this all together, for any $\delta > 0$, let $N_{\delta/\sqrt{n}}(\beta)\}$ be a $\delta/\sqrt{n}$ neighborhood of β. Then, we have that for any $\epsilon > 0$,

$$P\left(\sup_{b\in N_{\delta/\sqrt{n}}(\beta)}\left|\mathbb{G}_{\mathrm{n}}(xx'(\hat{f}(b)-\hat{f}(\beta)))\right|>h_n^{1/2}\epsilon\right) \tag{11}$$

$$\leq \frac{1}{\epsilon h_n^{1/2}}\,\mathbb{E}\left(\sup_{b\in N_{\delta/\sqrt{n}}(\beta)}|\,\mathbb{G}_{\mathrm{n}}(xx'(\hat{f}(b)-\hat{f}(\beta)))|\right) \tag{12}$$

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

Lemma 3. *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, Chetverikov, and Kato (2014):

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

where in the notation of Corollary 5.1 of Chernozhukov et al. (2014), we can choose $\sigma^2=O(1/\sqrt{n})$ by (10). This means that when $h_n=o(\log n/\sqrt{n})$, we obtain the desired result:

$$\frac{1}{\sqrt{n}}\,\mathbb{G}_{\mathrm{n}}(xx'(\hat{f}(b)-\hat{f}(\beta)))=o_P(1/\sqrt{nh_n}) \tag{13}$$

□