Diagnostic Testing of Finite Moment Conditions for the Consistency and Root-N Asymptotic Normality of the GMM and M Estimators*

Yuya Sasaki† and Yulong Wang‡

Abstract

Common econometric analyses based on point estimates, standard errors, and confidence intervals presume the consistency and the root-n asymptotic normality of the GMM or M estimators. However, their key assumption that data entail finite moments may not be always satisfied in applications. This paper proposes a method of diagnostic testing for these key assumptions. Applying the proposed test to the market share data from the Dominick's Finer Foods retail chain, we find that a common *ad hoc* procedure to deal with zero market shares in analysis of differentiated products markets results in a failure of these key assumptions.

Keywords: diagnostic test, consistency, root-n asymptotic normality

*First arXiv version: June 3, 2020. We thank numerous researchers and seminar participants at Durham University and University of Southern California for very useful comments. We thank James M. Kilts Center, University of Chicago Booth School of Business for allowing us to use this data set. All the remaining errors are ours.

†Associate professor of economics, Vanderbilt University. Email: yuya.sasaki@vanderbilt.edu

‡Assistant professor of economics, Syracuse University. Email: ywang402@maxwell.syr.edu.

1 Introduction

In empirical economic analysis, it is a common practice to drop outliers, whose definition can however be ambiguous. If the underlying distribution is normal, then deleting the observations that are larger than few standard deviations away from the sample mean seems reasonable; but such a practice could lead to a substantially biased estimate if the underlying distribution has a heavy tail or does not even have a finite mean. The classical confidence interval based on the t-statistic can also be misleading. Technically, the non-existence of outliers is defined as the finite first and second moment conditions on the score, which are required for consistent estimation and a valid t-statistic, respectively. These finite moment conditions are often taken for granted, but are not necessarily plausibly satisfied in empirical studies. Motivated by this concern, this article proposes a method of diagnostic testing for these finite moment conditions.[1]

To illustrate our method, let $A_i = A(D_i; \theta_0)$ denote the score function – e.g., $A_i = X_i U_i = X_i(Y_i - X_i^\intercal \theta_0)$ in case of the linear regression – for the estimator evaluated at the i-th observation D_i and the true parameter vector θ_0. To test the finite first moment condition $\mathbb{E}[|A_i|] < \infty$, we essentially require that the distribution of $|A_i|$ decays to zero at a power law, i.e.,

$$\lim_{a \to \infty} \frac{\mathbb{P}(|A_i| > a)}{a^{-1/\xi}} = 1,$$

where ξ is called the tail index that characterizes the decaying rate. Such a tail approximation is satisfied by many commonly used distributions - see Sections 4–5 for more discussions. Under this restriction, we have the elegant feature that the finite moment condition $\mathbb{E}[|A_i|] < \infty$ is equivalent to $\xi < 1$. Then, we propose a test of this equivalent condition $\xi < 1$ that has nontrivial asymptotic power in large samples.

More specifically, consider a test of finite r-th moment of the score, e.g., $r = 1$ (respectively, $r = 2$) for a test of the consistency (respectively, the root-n asymptotic normality). First, sort

[1] A closely related to but different from this objective is the literature on testing non- and weak-identification (e.g., Wright, 2003; Stock and Yogo, 2005; Inoue and Rossi, 2011; Sanderson and Windmeijer, 2016; Lee, McCrary, Moreira, and Porter, 2020). See the related literature subsection ahead for detailed discussions.

the r-th power of the estimated absolute score $|A(D_i, \hat{\theta})|$ in the descending order, where $\hat{\theta}$ is some consistent estimator of θ_0. Second, pick the largest k of these order statistics and self-normalize them. We show that these self-normalized statistics asymptotically follow a known joint distribution up to the unknown tail index parameter ξ. A sub-unit (respectively, super-unit) value of this tail index parameter ξ indicates a finite (respectively, infinite) r-th moment of the score. Using these dichotomous characteristics, we construct a likelihood ratio test based on the limiting joint distribution of the self-normalized statistics. We establish a size control property of the proposed test under the null hypothesis of a finite r-th moment of the score, which is our main result.

Simulation studies support the theoretical result of the size control property. Applying the proposed method of testing to the widely used market share data from Dominick's Finer Foods retail chain, we find that the common *ad hoc* treatment of zero market shares by adding an infinitesimal positive value results in a failure of the consistency and the root-n asymptotic normality. This failure results from the fact that inclusion of the logarithms of these infinitesimal numbers (i.e., large negative values) induces a heavy-tailed distribution of the regression residuals for observations with *originally non-zero* market shares.

Relation to the Literature: We are not aware of any existing paper that develops a diagnostic test of the finite moment condition for the consistency or the root-n asymptotic normality of the GMM or M estimators, as we do in this paper. This is potentially due to the well-known 'impossibility' results by Bahadur and Savage (1956) and Romano (2004). Romano (2004) shows that it is impossible to construct a powerful test that controls size *uniformly* over all possible underlying distributions. We show that it is still possible to construct tests that control size *pointwisely* under any fixed underlying distribution with a regularly varying tail. Such a framework is still flexible enough to cover many commonly used distributions and is widely adopted in the existing literature to model a variety of empirical datasets. To name a few, Rozenfeld, Rybski, Gabaix, and Makse (2011) model the city-size data under this framework and empirically justify Zipf's law. Gabaix, Gopikrishnan, Plerou, and Stanley (2003) establish a novel theory to explain the empirical Pareto tails in the stock market. Barro and Jin (2011) characterize the macroeconomic disaster dataset with a Pareto tail under this framework and estimate the coefficient of risk aversion. Following the influential paper by Piketty and Saez (2003), there have been many papers that model and explain the Pareto tail in the income and wealth distributions. See Gabaix (2016) for a comprehensive review.

A different but closely related topic is a set of tools to test non- and weak-identification (e.g., Wright, 2003; Stock and Yogo, 2005; Inoue and Rossi, 2011; Sanderson and Windmeijer, 2016; Lee, McCrary, Moreira, and Porter, 2020). These are related to our framework on one hand because non- and weak-identification also results in a failure of the canonical consistency and the root-n asymptotic normality, and therefore these testing methods serve for related objectives. On the other hand, these are different from our framework because the non- and weak-identification concerns about non- and weak-invertibility of the expected gradient of the score,[2] whereas the issue of our concern is instead about the finiteness of moments of the score as the conditions for the WLLN and CLT. In this sense, the purposes of our method of test are different from those of the preceding methods of tests of non- and weak-identification, while they indeed play complementary roles.

Also related is the paper by Shao, Yu, and Yu (2001) that proposes a test of finite variance. On the one hand, our test of the root-n asymptotic normality is also based on the test of finite second moments, similarly to Shao, Yu, and Yu (2001). On the other hand, our objective of testing the asymptotic normality for the GMM and M estimators requires to take into account that the score is not directly observed in data, but has to be estimated via the GMM or M estimation. With these similarities and differences, our proposed method also contributes to this existing literature on testing finite moments by allowing for generated data.

For scalar locations and single equation models, an infinite first or second moment of the

[2]For general matrix rank tests, see e.g., Gill and Lewbel (1992); Cragg and Donald (1996, 1997); Robin and Smith (2000); Kleibergen and Paap (2006); Camba-Méndez and Kapetanios (2009); Al-Sadoon (2017).

score is often imputed to outliers. Edgeworth (1887) proposes to use the absolute loss instead of the square loss for a robust estimation of the equation parameters. This idea later extends and generalizes to other robust methods based on the check losses (Koenker and Bassett Jr, 1978) and the Huber loss (Huber, 1992). While we propose a test of the finite second moment of a norm of the score for the root-n asymptotic normality of GMM and M estimators in general, there are existing papers that establish limit distribution theories (which are not necessarily root-n or normal) without requiring the finite second moment condition in these frameworks (e.g., Davis and Resnick, 1985, 1986; Davis, Knight, and Liu, 1992; Hill and Prokhorov, 2016).

Finally, our method is based on recent developments in extreme value theory. We refer readers to De Haan and Ferreira (2006) for a very comprehensive review of this subject. In particular, our inference approach is based on the fixed-k asymptotics and takes advantage of the technique developed by Müller and Wang (2017) and Müller (2020). Müller and Wang (2017) construct fixed-k confidence intervals for the extreme quantile and tail conditional expectation, based on a random sample drawn from some unknown distribution. In comparison, we focus on the tail index and more importantly, allow for generated variables. The fixed-k approach is useful in practice, because the asymptotic size control is valid for any predetermined fixed number k, unlike traditional increasing-k approaches that require a sequence of changing tuning parameters as the sample size grows for which a sensible choice rule is difficult to obtain in small samples. The fixed-k approach also allows for robustness against errors in preliminary estimation – this type of robustness, benefiting from the fixed-tuning parameter setup, has been similarly explored in other contexts in the existing literature, e.g., the fixed-b asymptotic inference under heteroskedasticity and autocorrelation proposed by Kiefer and Vogelsang (2005) and the robust inference in kernel estimations proposed by Cattaneo, Crump, and Jansson (2014). In constructing our likelihood ratio test, we take advantage of the computational algorithm developed by Elliott, Müller, and Watson (2015).

Organization: The rest of this article is organized as follows. Section 2 sets up the econometric frameworks and Section 3 characterizes the testing problem. Section 4 introduces our test and previews its properties, followed by asymptotic derivations in Section 5. Section 6 contains simulation studies and Section 7 conducts a detailed study of the market shares data set. Section 8 concludes with proofs and omitted details relegated to the Supplementary Applendix.

2 Econometric Frameworks

We first introduce the general frameworks of the GMM and M estimators under which we propose the new test. A concrete empirical example will follow after the presentation of the general frameworks.

2.1 GMM and M Estimators

M-Estimation: Consider the class of estimators defined by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{Q}_n(\theta),$$

where the criterion function $\hat{Q}_n$ takes the form of $\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^n g_i(\theta)$. Under regularity conditions for this class (cf. Newey and McFadden, 1994, Section 3.2), the influence function representation takes the form of $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\hat{H}_n(\theta_0)^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i'(\theta_0) + o_p(1)$, where $\hat{H}_n(\theta) = n^{-1} \sum_{i=1}^n D_\theta^2 g_i(\theta)$ and $g_i'(\theta) = \nabla_\theta g_i(\theta)$. The consistency of $\hat{\theta}$ (via the weak law of large numbers) requires $\mathbb{E}\left[\|g_i'(\theta_0)\|\right] < \infty$. Likewise, the asymptotic normality of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ (via multivariate Lindeberg-Lévy CLT) requires $\mathbb{E}\left[\|g_i'(\theta_0)\|^2\right] < \infty$. Common examples include the following two classes of estimators:

1. (OLS) $g_i'(\theta) = X_i\left(Y_i - X_i^\intercal \theta\right) \Rightarrow A_i^r(\theta) \equiv \|g_i'(\theta)\|^r = \{(Y_i - X_i^\intercal \theta)^\intercal X_i^\intercal X_i \left(Y_i - X_i^\intercal \theta\right)\}^{r/2}$

2. (QMLE) $g_i'(\theta) = \nabla_\theta \ell(Y_i, X_i; \theta) \Rightarrow A_i^r(\theta) \equiv \|g_i'(\theta)\|^r = \{\nabla_\theta \ell(Y_i, X_i; \theta)^\intercal \nabla_\theta \ell(Y_i, X_i; \theta)\}^{r/2}$

In this paper, we propose tests of the null hypothesis: $\mathbb{E}\left[A_i^1(\theta_0)\right] < \infty$, the condition that is required for establishing the consistency of $\hat{\theta}$; and the null hypothesis: $\mathbb{E}\left[A_i^2(\theta_0)\right] < \infty$, the condition that is required for establishing the asymptotic normality of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$.

Remark 1. *Due to the related structures between M and Z estimators, our framework also applies to Z estimators.*

GMM: Next, consider the class of estimators defined by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{Q}_n(\theta),$$

where the criterion function $\hat{Q}_n$ takes the form of $\hat{Q}_n(\theta) = \left[n^{-1} \sum_{i=1}^n g_i(\theta)\right]^\intercal \hat{W} \left[n^{-1} \sum_{i=1}^n g_i(\theta)\right].$ Under regularity conditions for this class (cf. Newey and McFadden, 1994, Section 3.3), the influence function representation takes the form of $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -(\hat{G}_n(\hat{\theta})^\intercal \hat{W}_n \hat{G}_n(\theta_0))^{-1} \hat{G}_n(\hat{\theta})^\intercal \hat{W}_n$ $n^{-1/2} \sum_{i=1}^n g_i(\theta_0) + o_p(1)$, where $\hat{G}_n(\theta) = n^{-1} \sum_{i=1}^n \nabla_\theta g_i(\theta)$ and $\hat{W}_n \xrightarrow{p} W_0$. The consistency of $\hat{\theta}$ (via the weak law of large numbers) requires $\mathbb{E}\left[\|g_i(\theta_0)\|\right] < \infty$. Likewise, the asymptotic normality of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ (via multivariate Lindeberg-Lévy CLT) requires $\mathbb{E}\left[\|g_i(\theta_0)\|^2\right] < \infty$. A common example is:

3 (2SLS) $g_i(\theta) = Z_i\left(Y_i - X_i^\intercal \theta\right) \Rightarrow A_i^r(\theta) \equiv \|g_i(\theta)\|^r = \left\{(Y_i - X_i^\intercal \theta)^\intercal Z_i^\intercal Z_i \left(Y_i - X_i^\intercal \theta\right)\right\}^{r/2}$

Similarly to the M-estimation case, we propose tests of the null hypothesis: $\mathbb{E}\left[A_i^1(\theta_0)\right] < \infty$, which is required for establishing the consistency of $\hat{\theta}$; and the null hypothesis: $\mathbb{E}\left[A_i^2(\theta_0)\right] < \infty$, which is required for establishing the asymptotic normality of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$.

2.2 Example: Demand Analysis in Differentiated Products Markets

In applications, $A_i^r(\theta_0)$ may not have a finite moment in the presence of outliers in the dependent variable. Outliers may be innate in data in some applications. In other applications, outliers may be produced as artifacts of *ad hoc* procedures taken by researchers. In the empirical application to be presented in Section 7 ahead, we highlight this point in the context of the demand analysis under the following setup.

Example 1 (Demand Analysis)**.** *Demand estimation with market share data*[3] *in the logit case is based on GMM with the moment function defined by*

$$g_{jt}(\theta) = Z_{jt}\left(\ln(S_{jt}) - \ln(S_{0t}) - P_{jt}\theta_1 - X_{jt}^{\intercal}\theta_{-1}\right),$$

where j *indexes products,* t *indexes markets,* S_{jt} *denotes the share of product* j *in market* t*,* P_{jt} *denotes the price,* X_{jt} *denotes product characteristics, and* Z_{jt} *denotes instruments. In this setting,*

$$A_{jt}^r(\theta) = \{(\ln(S_{jt}) - \ln(S_{0t}) - P_{jt}\theta_1 - X_{it}^{\intercal}\theta_{-1})^{\intercal} Z_{jt}^{\intercal} Z_{jt} (\ln(S_{jt}) - \ln(S_{0t}) - P_{jt}\theta_1 - X_{it}^{\intercal}\theta_{-1})\}^{r/2},$$

evaluated at $\theta = \theta_0$*, may not have a finite moment when* $\ln(S_{jt})$ *has a heavy-tailed distribution. Furthermore, in the presence of zero market shares, there is a common ad hoc practice of replacing zero shares by infinitesimal shares in order to avoid the logarithm of zeros, but only to artificially produce a heavy-tailed distribution of regression residuals for observations with "originally non-zero" market shares. Such a practice can result in a failure of the consistency and the root-n asymptotic normality – see our empirical applications in Section 7.*

3 Hypothesis Testing Problem

This section establishes the equivalence between the event of finite moments and the event of sub-unit values of the tail index.

Consider the distribution function F of a non-negative random variable, examples of which are the random variables $A_i^r(\theta_0)$ that were introduced in Section 2. As a preliminary step before constructing a test, we show that whether its moment is finite is fully determined by

[3]This framework is drawn from the literature on the estimation of demand for differentiated products (Berry, 1994; Berry, Levinsohn, and Pakes, 1995). See Ackerberg, Benkard, Berry, and Pakes (2007) for a survey.

the right-tail behavior of F. To this end, we introduce our key assumption on F. We say that a distribution F is within the domain of attraction (DoA) of the extreme value distribution, denoted by $F \in \mathcal{D}(G_\xi)$, if there exist sequences of constants a_n and b_n such that for every v,

$$\lim_{n\to\infty} F^n(a_n v + b_n) = G_\xi(v)$$

holds, where the function G_ξ is defined by

$$G_\xi(v) = \begin{cases} \exp\left(-(1+\xi v)^{-1/\xi}\right) & 1+\xi v > 0, \xi \neq 0 \\ \exp\left(-e^{-v}\right) & v \in \mathbb{R}, \xi = 0, \end{cases}$$

and is referred to as the generalized extreme value distribution.

This condition, characterizing the tail shape of the underlying distribution, is mild and satisfied by many commonly used distributions. In particular, the case with $\xi > 0$ equivalently covers regularly varying (RV) distributions such that

$$1 - F(a) = a^{-1/\xi}\mathcal{L}(a) \tag{3.1}$$

for some function $\mathcal{L}(\cdot)$ that satisfies $\mathcal{L}(a) \to 1$ as $a \to \infty$. Leading examples include Pareto, Student-t, and F distributions. The case with $\xi \leq 0$ covers thin tailed distributions that have a finite r-th moment for any $r > 0$, including, for example, the Gaussian family and distributions with bounded supports. See De Haan and Ferreira (2006, Ch.1) for a comprehensive review. Since the finite moments are readily satisfied in this case, we focus on the case of $\xi > 0$ in the rest of the paper. We will interchangeably refer to DoA and RV as our key assumption.

The following lemma formally characterizes the finiteness of the moment in terms of the sub-unit values of ξ. A proof can be found in Appendix A.1.

Lemma 1 (Characterization)**.** *For a generic non-negative random variable A with distribution F_A satisfying (3.1), we have that*

$$\mathbb{E}[A] < \infty \text{ if } \xi < 1;$$

$$\mathbb{E}[A] = \infty \text{ if } \xi > 1.$$

In addition, suppose $\mathcal{L}(\cdot)$ is uniformly bounded below from zero. Then $\mathbb{E}[A] = \infty$ if $\xi = 1$.

A couple of remarks are in order. First, the results with $\xi < 1$ and $\xi > 1$ have been stated in Mikosch (1999) with proofs following from Karamata's theorem (e.g., Resnick, 2007, Theorem 2.1). We present them here mainly for completeness. Second, the case with $\xi = 1$ is more complicated since the distribution in such case may have either infinite or finite first moment, if no further restriction is imposed. To explicitly characterize this case, we impose the additional condition that $\mathcal{L}(\cdot)$ is uniformly bounded below from zero. This is mild and satisfied by many commonly used distributions. For example, if the underlying distribution is exactly Pareto, we have $\mathcal{L}(\cdot) = 1$. If the underlying distributio is Student-t, we have that $\mathcal{L}(a) = c(1 + da^{-2} + o(a^{-2})$ as $a \to \infty$ for some constants $c > 0$ and $d \neq 0$.

In summary, Lemma 1 implies that, given the class of distributions that satisfy the DoA condition, the tail index ξ cannot exceed one for any distribution that has a finite moment. Then, we are ready to state our competing hypotheses as follows

$$H_0 : \xi \in (0, 1) \text{ against } H_1 : \xi \in [1, \bar{\xi}], \tag{3.2}$$

where $\bar{\xi}$ is the upper bound of the parameter space that includes all empirically relevant values of ξ. We set $\bar{\xi} = 2$ in later sections, which can be easily extended. Note that the null space shares the boundary of the alternative space at the point $\xi = 1$. At such boundary point, our proposed test is expected to have equal size and power, which is presented later in Figure 1.

4 The Test

In order to test the hypotheses in (3.2), we would like to observe large values of $A_i^r(\theta_0)$, which are not available since θ_0 is unknown. We make use of a consistent estimator $\hat{\theta}$ of θ_0. Let $A_{(1)}^r(\hat{\theta}) \geq A_{(2)}^r(\hat{\theta}) \geq \ldots \geq A_{(n)}^r(\hat{\theta})$ denote the order statistics by sorting $\{A_i^r(\hat{\theta})\}_{i=1}^n$ in the descending order. For a pre-determined integer $k \geq 3$, collect the largest k order statistics as

$$\mathbf{A}^r\left(\hat{\theta}\right) = \left[A_{(1)}^r(\hat{\theta}), A_{(2)}^r(\hat{\theta}), \ldots, A_{(k)}^r(\hat{\theta})\right]^\intercal. \tag{4.1}$$

By extreme value theory again, the joint distribution of the largest order statistics asymptotically approaches a well-defined parametric joint distribution that is fully characterized by the location, the scale, and the scalar parameter ξ. Therefore, if we conduct location and scale normalization by considering the statistics

$$\mathbf{A}_*^r\left(\hat{\theta}\right) = \frac{\mathbf{A}^r\left(\hat{\theta}\right) - A_{(k)}^r\left(\hat{\theta}\right)}{A_{(1)}^r\left(\hat{\theta}\right) - A_{(k)}^r\left(\hat{\theta}\right)},$$

where for any generic $k \times 1$ vector $\mathbf{A} = (A_1, \ldots, A_k)^\intercal$ and scalars $a \neq 0$ and b, the notation $a^{-1}(\mathbf{A} - b)$ is understood componentwisely as $(a^{-1}(A_1 - b), \ldots, a^{-1}(A_k - b))^\intercal$, then $\mathbf{A}_*^r(\hat{\theta})$ converges in distribution to a limiting random vector $\mathbf{V}_*$, whose density $f_{\mathbf{V}_*}$ is fully characterized by ξ and is invariant to location, scale, and the order r – see (5.2) ahead for the formula of $f_{\mathbf{V}_*}$. By construction, the first and the last elements of $\mathbf{A}_*^r(\hat{\theta})$ (and $\mathbf{V}_*$) are one and zero, respectively. The estimation error in $\hat{\theta}$ is asymptotically negligible since it is of a smaller order of magnitude than the largest order statistics of $A_i^r(\theta_0)$ under the null hypothesis – see Section 5 for details.

Now, the limiting testing problem has become straightforward: we construct a test based on a random draw of $\mathbf{V}_*$ from its parametric density $f_{\mathbf{V}_*}$ about the only unknown scalar parameter ξ. When the null and alternative hypotheses are both simple, the optimal solution is known to be the Neyman-Pearson test, where large values of the likelihood ratio statistic reject the null hypothesis. Therefore, we transform the null and alternative hypotheses of (3.2) into simple ones by considering weighted average likelihoods, and our proposed test rejects the null hypothesis that $A_i^r(\theta_0)$ has a finite moment if

$$\frac{\int f_{\mathbf{V}_*}\left(\mathbf{A}_*^r(\hat{\theta}); \xi\right) dW(\xi)}{\int f_{\mathbf{V}_*}\left(\mathbf{A}_*^r(\hat{\theta}); \xi\right) d\Lambda(\xi)} > \text{cv},$$

where cv denotes the critical value, $W(\cdot)$ denotes a weight chosen by the econometrician to reflect the importance of rejecting different alternatives, and $\Lambda(\cdot)$ is some pre-determined weight defined on the null space. More details about this test are presented in the following section.

We provide some heuristic discussions of the *asymptotic* property of the new test - formal discussions will follow in Section 5. Figure 1 plots oracle rejection probabilities of the test with $\mathbf{V}_*$ generated from the limiting distribution $f_{\mathbf{V}_*}$ with various values of ξ and the nominal size of 0.05. The plots are based on 10000 simulation draws. (This is a power prescription and is different from Monte Carlo simulation studies. Full-blown Monte Carlo simulation studies with concrete econometric models and small samples will be conducted and presented in Section 6 to evaluate the *finite sample* performance.) Observe that the rejection probabilities for $\xi \in (0,1)$ are dominated by the nominal size, 0.05. In other words, the test has a size control property for all distributions with tail index less than one.

The nontrivial power of our proposed test is in fact not at odds with the result of Romano (2004, Example 1), which is often viewed as an 'impossibility' result regarding tests of finite moments. Specifically, Romano (2004, Theorem 1) shows that $\sup_{P\in\mathcal{P}_1}\mathbb{E}_P[\phi(D)] \leq \sup_{P\in\mathcal{P}_0}\mathbb{E}_P[\phi(D)]$ holds for any test $\phi(\cdot)$, where $\mathcal{P}_0$ and $\mathcal{P}_1$ denote the sets of distributions compatible with the null and alternative hypotheses, respectively. His result implies that even the maximum power in the alternative space cannot exceed the uniform size if we allow the data generating process to change with the sample size. To illustrate this point, consider the Pareto distribution $F_1(y) = 1 - y^{-1/2}$ that belongs to $\mathcal{P}_1$. We can construct the sequence of distributions $\{F_{0m}(y)\}_{m=1}^{\infty}$ such that $F_{0m}(y)$ is identical to $F_1(y)$ if y is less than the $1-1/m$ quantile of F_1 and $1 - y^{-2}$ otherwise. Note that each F_{0m} belongs to $\mathcal{P}_0$. By construction, F_{0m} and F_1 differ from each other only at the very tail, and hence the extreme order statistics cannot distinguish F_1 from F_{0m} if we allow m to increase with n sufficiently fast. On the other hand, many empirical studies aforementioned in the introduction implicitly adopt the framework that the data are generated from some fixed underlying distributions (m is fixed and n is increasing). Under this framwork, the largest k order statistics will stem almost surely from the tail part of the distribution, and hence they will nontrivially inform us of the true tail behavior of the distribution asymptotically. This feature allows for the power to exceed the size of the

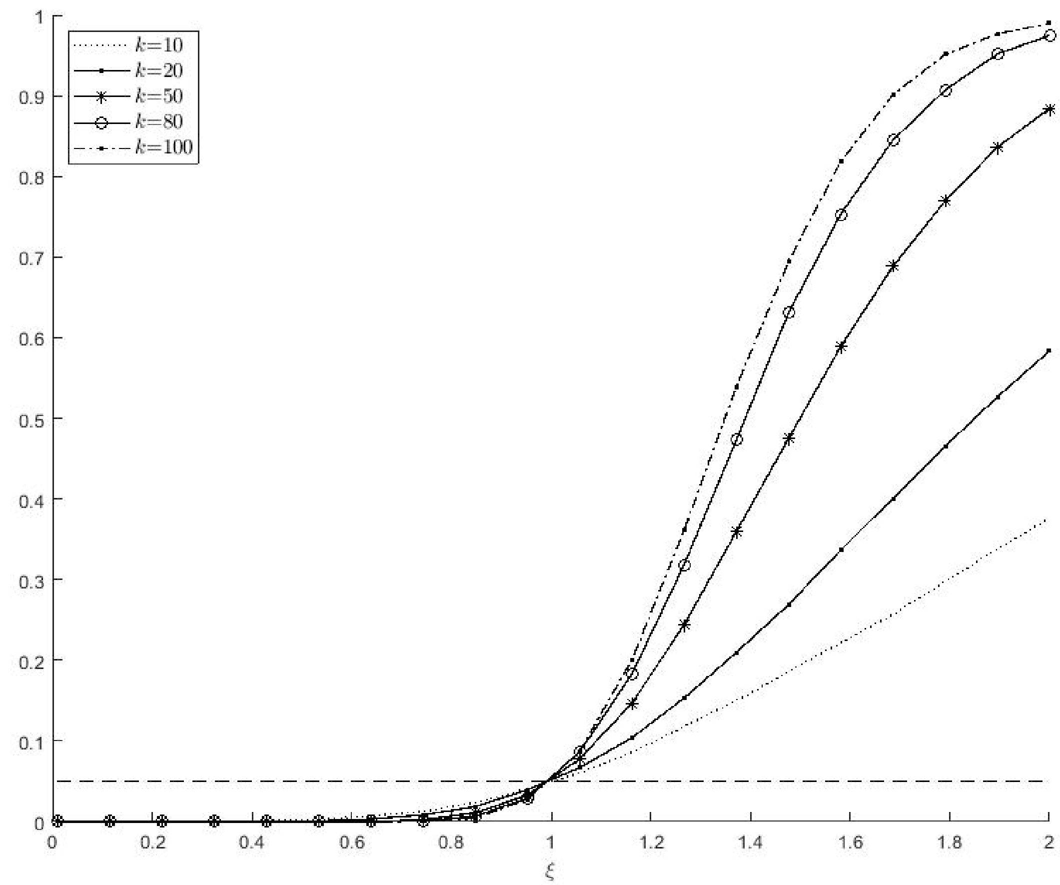


Figure 1: Rejection probabilities of the test with $\mathbf{V}_*$ generated from $f_{\mathbf{V}_*}$ with $\xi \in [0.01, 2]$ and the nominal size of 0.05. The plots are based on 10000 simulations.

test, as demonstrated in Figure 1.

5 Asymptotic Theory

We now present a formal theory to guarantee that our proposed test works in large samples. To this end, we first introduce additional notations and some definitions. Denote by D_i the i-th observation so that we can write $A_i^r(\theta) = A^r(\theta; D_i)$. For example, $D_i = (X_i^\intercal, Y_i)^\intercal$ in the context of the OLS, and $D_i = (X_i^\intercal, Z_i^\intercal, Y_i)^\intercal$ in the context of the 2SLS presented in Section 2.1. Let $F_{A^r(\theta)}$ denote the cumulative distribution function (CDF) of $A_i^r(\theta)$ and θ_0 denote the (pseudo-) true value of θ. Let $B_{\eta_n}(\theta_0)$ denote an open ball centered at θ_0 with radius $\eta_n \to 0$. Let $f_{A^r(\theta_0)}$ and $Q_{A^r(\theta_0)}$ be the probability density function (PDF) and quantile function of $A_i^r(\theta_0)$, respectively. With these notations and definitions, we impose the following regularity conditions to prove the size control property of our proposed test.

Condition 1. *The following conditions are satisfied.*

(i) D_i *is i.i.d. from some underlying distribution that does not change with* n.

(ii) $F_{A^r(\theta_0)}$ *satisfies (3.1) with* $\mathcal{L}(\cdot)$ *bounded below from zero.*

(iii) $\hat{\theta} - \theta_0 = o_p(1)$. *For some* $\eta_n \to 0$, $\sup_i \sup_{\theta \in B_{\eta_n}(\theta_0)} \left|\left|\frac{\partial A_i^r(\theta)}{\partial \theta}\right|\right| = O_p\left(n^{\xi}\right)$;

Condition 1 (i) requires random sampling from some fixed population distribution. It rules out the trivial case where the location and the scale of the data diverge with the sample size. As a consequence, the existence of the moment only depends on the tail heaviness of the underlying distribution, which should remain invariant to location and scale shifts. Condition 1 (ii) requires that the distribution of $A^r(\theta_0)$ falls in the domain of attraction of the extreme value distribution, which bridges the finiteness of moments and the tail heaviness.

The first part of Condition 1 (iii) requires that the estimator $\hat{\theta}$ is consistent for θ_0. Note that this in general only requires the identification and finite first moments of the score. For the case of testing the finite first moment condition for consistency (i.e., the case of setting $r = 1$), this assumption is satisfied under the null hypothesis in (3.2). For the case of testing the finite second moment condition for root-n asymptotic normality (i.e., the case of setting $r = 2$), this assumption is satisfied even under a range of alternative hypotheses as well as under the null hypothesis. In any case, the fact that this consistency condition is satisfied under the null hypothesis allows us to establish a size control property based on this condition.

The second part of Condition 1 (iii) requires that the gradient of $A_i^r(\theta)$ grows not too fast as the sample size increases. Since this last piece of the condition is a high-level statement, it will be useful to consider stronger lower level sufficient conditions in a specific example. We do so in the context of the GMM estimator, which is implemented in our application.

It is worth mentioning that the consistency of $\hat{\theta}$ is sufficient but not necessary for the size control. It can be relaxed provided a tigher bound on the magnitude of gradient. Consider OLS for example, where the gradient becomes $X_i X_i^\intercal$. If X_i has a bounded support, the gradient is then uniformly bounded over i and then Lemma 2 holds as long as the parameter space of θ_0 is also bounded.

Discussion of Condition 1 (iii) – Case of GMM: Consider the case of setting $r = 2$ for testing the root-n asymptotic normality. Recall that we define $A_i^2(\theta) = (Y_i - X_i^\intercal\theta)^\intercal Z_i^\intercal Z_i (Y_i - X_i^\intercal\theta)$. Thus,

$$\begin{aligned} \frac{\partial A_i^2(\theta)}{\partial\theta} &= -2X_i Z_i^\intercal Z_i (Y_i - X_i^\intercal\theta) \\ &= -2X_i Z_i^\intercal Z_i u_i + 2X_i Z_i^\intercal Z_i X_i^\intercal \left(\hat{\theta} - \theta\right). \end{aligned}$$

The triangle inequality and Cauchy-Schwartz inequality yield

$$\begin{aligned} \sup_i \sup_{\theta\in B_{\eta n}(\theta_0)} \left|\left|\frac{\partial A_i^2(\theta)}{\partial\theta}\right|\right| &\leq 2\sup_i ||X_i Z_i^\intercal Z_i u_i|| + 2\sup_i ||X_i Z_i^\intercal Z_i X_i^\intercal|| \cdot \sup_{\theta\in B_{\eta n}(\theta_0)} \left|\left|\hat{\theta} - \theta\right|\right| \\ &\leq 2\sup_i ||X_i Z_i^\intercal|| \left(\sup_i \left|\left|A_i^2(\theta_0)\right|\right|\right)^{1/2} + 2\left(\sup_i ||X_i Z_i^\intercal||\right)^2 \cdot \sup_{\theta\in B_{\eta n}(\theta_0)} \left|\left|\hat{\theta} - \theta\right|\right|. \quad (5.1) \end{aligned}$$

Condition 1 (ii) implies that $\sup_i ||A_i^2(\theta_0)|| = O_p(n^\xi)$ so that $(\sup_i ||A_i^2(\theta_0)||)^{1/2}$ is of a smaller order than $\sup_i ||A_i^2(\theta_0)||$. Therefore, a sufficient condition for Condition 1 (iii) is that $\sup_i ||X_i Z_i^\intercal||$ is of a smaller order than $(\sup_i ||A_i^2(\theta_0)||)^{1/2}$. An even stronger sufficient condition for this sufficient condition is that X_i and Z_i have bounded supports, which is satisfied by the typical applications in the demand analysis, in particular the one that we consider in our empirical application in Section 7. □

We now proceed to derive the asymptotic property of the proposed test. The following lemma shows that the largest order statistics as in (4.1) asymptotically follow the joint extreme value distribution.

Lemma 2. *Under Condition 1, there exist sequences of constants a_n and b_n depending on $F_{A^r(\theta_0)}$ such that, for any fixed k,*

$$\frac{\mathbf{A}^r\left(\hat{\theta}\right)-b_n}{a_n}\xrightarrow{d}\mathbf{V}\equiv\left(V_1,...,V_k\right)^{\intercal},$$

where $\mathbf{V}$ is jointly distributed with the density given by $f_{\mathbf{V}|\xi}(v_1,...,v_k)=G_\xi(v_k)\prod_{i=1}^k g_\xi(v_i)/G_\xi(v_i)$ on $v_k\leq v_{k-1}\leq\ldots\leq v_1$, and $g_\xi(v)=\partial G_\xi(v)/\partial v$.

A proof is provided in Appendix A.2.

Since a_n and b_n are unknown, we would like to eliminate them in constructing feasible test statistics. We do so by constructing the self-normalized statistic

$$\mathbf{A}_*^r\left(\hat{\theta}\right)=\left(1,\frac{A_{(2)}^r\left(\hat{\theta}\right)-A_{(k)}^r\left(\hat{\theta}\right)}{A_{(1)}^r\left(\hat{\theta}\right)-A_{(k)}^r\left(\hat{\theta}\right)},\ldots,\frac{A_{(k-1)}^r\left(\hat{\theta}\right)-A_{(k)}^r\left(\hat{\theta}\right)}{A_{(1)}^r\left(\hat{\theta}\right)-A_{(k)}^r\left(\hat{\theta}\right)},0\right)^{\intercal}.$$

By the continuous mapping theorem, change of variables, and Lemma 2, we obtain

$$\mathbf{A}_*^r\left(\hat{\theta}\right)\xrightarrow{d}\mathbf{V}_*\equiv\frac{\mathbf{V}-V_k}{V_1-V_k},$$

where the density function $f_{\mathbf{V}_*}$ of the limit observation $\mathbf{V}_*$ is given by

$$f_{\mathbf{V}_*}\left(\mathbf{v}_*;\xi\right)=\Gamma\left(k\right)\int_0^\infty s^{k-2}\exp\left(\left(-1-1/\xi\right)\left(\log(1+\xi s)+\sum_{i=2}^{k-1}\log\left(1+\xi v_{*i}s\right)\right)\right)ds,\quad(5.2)$$

and v_{*i} denotes the i-th component of $\mathbf{v}_*$. With this density function, we construct the likelihood ratio test

$$\varphi\left(\mathbf{V}_*\right)=\mathbf{1}\left[\frac{\int_1^2 f_{\mathbf{V}_*}\left(\mathbf{V}_*;\xi\right)dW\left(\xi\right)}{\int_0^1 f_{\mathbf{V}_*}\left(\mathbf{V}_*;\xi\right)d\Lambda\left(\xi\right)}>\mathrm{cv}\right],\quad(5.3)$$

where $W\left(\cdot\right)$ denotes a weight chosen by the econometrician and $\Lambda\left(\cdot\right)$ is some pre-determined weight that transforms the composite null space into a simple one. The critical values are

k	10	20	30	40	50	60	70	80	90	100
cv	2.15	2.57	2.65	2.59	2.45	2.40	2.27	2.10	2.22	1.98
k	150	200	250	300	350	400	450	500	1000	2000
cv	1.51	1.34	1.14	1.12	1.03	0.92	0.87	0.80	0.59	0.42

Table 1: 5% critical values of the test (5.4). Based on 10000 simulation draws.

determined by simulations. In particular, $W(\xi)$ reflects the weight attached to the rejection probability under the value of ξ in the alternative space. Since ξ is unknown, our likelihood ratio test is essentially designed to maximize the local average power with respect to W (cf. Andrews and Ploberger, 1994). We choose the uniform weight in later sections mainly for implementational simplicity, which can be easily changed.

It remains to determine the weight Λ, which is referred to as the least favorable distribution (e.g., Lehmann and Romano, 2005, Ch.3.8) that guarantees the size control over the null hypothesis. To this end, we adapt the generic algorithm developed by Elliott, Müller, and Watson (2015) to numerically construct Λ - see Appendix B for details. It turns out that Λ allocates all the weights to the point $\xi = 1$, suggesting that the least favorable distribution is simply the point mass on the boundary of the null and the alternative spaces. Using the uniform weight for $W(\cdot)$, the test (5.3) then reduces to

$$\varphi(\mathbf{V}_*) = \mathbf{1}\left[\frac{\int_1^2 f_{\mathbf{V}_*}(\mathbf{V}_*;\xi)\, d\xi}{f_{\mathbf{V}_*}(\mathbf{V}_*;1)} > \mathrm{cv}\right], \tag{5.4}$$

which is implemented in later sections. The critical values at the 5% nominal level are presented across alternative values of k in Table 1.

The following theorem establishes the asymptotic size control of our test (5.3).

Theorem 1. *Suppose Condition 1 holds. Then for any fixed k, under H_0 in (3.2)*

$$\lim_{n\to\infty} \mathbb{E}\left[\varphi\left(\mathbf{A}_*^r(\hat{\theta})\right)\right] \leq \alpha.$$

A proof is provided in Appendix A.3.

We close this section with two remarks. First, the fixed-k asymptotic design leads to the desired size control property as stated in the theorem. This feature provides a practical advantage because a researcher does not *ex ante* know the true distribution under the composite null hypothesis.

Second, our method is robust to estimation errors and hence allows for generated observations, which have not been discussed in the existing literature about the fixed-k design (e.g., Müller and Wang, 2017). Such robustness comes from the fact that the estimation error, $\hat{\theta} - \theta_0$ is $o_p(1)$ and hence to be dominated by the infeasible largest order statistics of $\{A_i^r(\theta_0)\}$.

6 Simulation Studies

Using Monte Carlo simulations, we demonstrate that the proposed test has the claimed size control property. We consider two of the most popular econometric models, namely the linear regression model and the linear IV model, for data generating designs. For each of these two designs, we consider the test of the consistency (by setting $r = 1$) and the test of the root-n asymptotic normality (by setting $r = 2$).

6.1 Simulation Setup

First, consider the linear regression model:

$$Y_i = \theta_1 + \theta_2 X_i + U_i,$$

where $\theta = (\theta_1, \theta_2)^\intercal = (1, 1)^\intercal$. The independent variable is generated according to $X_i \sim N(0, 1)$. The error U_i is independent from X_i and generated from the symmetric generalized Pareto distribution with tail index ξ_U that satisfies $1 - F_U(u) = (1 + \xi_U u)^{-1/\xi_U}/2$ for $u \geq 0$. We vary the value of ξ_U across sets of simulations. The test is based on the r-th moment of the score: $A_i^r(\theta) = (1 + X_i^2)^{r/2} \cdot |Y_i - \theta_1 - \theta_2 X_i|^r$ for $r = 1, 2$. Since we do not know θ_0, we replace θ_0 by the OLS $\hat{\theta}$. We thus use k order statistics of

$$A_i^r(\hat{\theta}) = (1 + X_i^2)^{r/2} \cdot |Y_i - \hat{\theta}_1 - \hat{\theta}_2 X_i|^r$$

to construct our test, following the procedure outlined in Section 4.

Second, consider the linear IV model:

$$\begin{aligned} Y_i =& \theta_1 + \theta_2 X_i + U_i + V_i \\ X_i =& \pi_1 + \pi_2 Z_i + R_i \end{aligned}$$

where $\theta = (\theta_1, \theta_2)^\intercal = (1, 1)^\intercal$ and $\pi = (\pi_1, \pi_2)^\intercal = (1, 1)^\intercal$. The instrument is generated according to $Z_i \sim N(0, 1)$ independently of the tri-variate error components $(U_i, V_i, R_i)^\intercal$. The heavy-tailed part of the error U_i is generated in the same way as above, independently from $(V_i, R_i)^\intercal$. We vary the value of ξ_U across sets of simulations. The endogenous part of the error components $(V_i, R_i)^\intercal$ is generated according to $(V_i, R_i)^\intercal \sim N(\vec{0}, \Sigma)$ where $\Sigma = (1, 0.5; 0.5, 1)$. The test is based on the r-th moment of the score: $A_i^r(\theta) = (1 + Z_i^2)^{r/2} \cdot |Y_i - \theta_1 - \theta_2 X_i|^r$ for $r = 1, 2$. Since we do not know θ_0, we replace θ_0 by the IV estimator $\hat{\theta}$. We thus use k order statistics of

$$A_i^r(\hat{\theta}) = (1 + Z_i^2)^{r/2} \cdot |Y_i - \hat{\theta}_1 - \hat{\theta}_2 X_i|^r$$

to construct our test, following the procedure outlined in Section 4.

For each of the linear regression model and the linear IV model introduced above, we experiment with sample sizes of $n = 10^4$, 10^5 and 10^6, which are similar to the sample size that we actually encounter in our empirical application in Section 7. For testing the finite first moment condition, we experiment with the tail index values of $\xi_U = 0.19$, 0.39, 0.59, 0.79, 0.99, 1.19, 1.39, 1.59, 1.79 and 1.99. Note that $\xi_U \in \{0.19, 0.39, 0.59, 0.79, 0.99\}$ satisfy the condition for the consistency, but $\xi_U \in \{1.19, 1.39, 1.59, 1.79, 1.99\}$ fail to satisfy it. For testing the finite second moment, we experiment with the tail index values of $\xi_U = 0.09$, 0.19, 0.29, 0.39, 0.49, 0.59, 0.69, 0.79, 0.89 and 0.99. Note that $\xi_U \in \{0.09, 0.19, 0.29, 0.39, 0.49\}$ satisfy the condition for the asymptotic normality, but $\xi_U \in \{0.59, 0.69, 0.79, 0.89, 0.99\}$ fail to satisfy it. We also experiment with various numbers $k = 50$, 100, and 200 of order statistics for construction of the test. Each set of simulations consists of 5000 Monte Carlo draws.

6.2 Simulation Results

Table 2 shows Monte Carlo simulation results of testing the finite first moment condition in (A) the linear regression model and (B) the linear IV model. In both of the two panels, (A) and (B), we can see that the simulated rejection probabilities are dominated by the nominal size 0.05 for all of $\xi_U \in \{0.19, 0.39, 0.59, 0.79\}$ in the null region, and those are approximately the same as the nominal size 0.05 near the boundary, i.e., $\xi_U = 0.99$, of the null region. These results support the size control property of the test that is established in Theorem 1. The condition for the consistency holds for any of $\xi_U < 1$, but a researcher does not *ex ante* know or does not want to pick which exact value ξ_U for a specific application. For this reason, this size control property is important in practice.

Table 3 shows Monte Carlo simulation results of testing the finite second moment condition in (A) the linear regression model and (B) the linear IV model. The findings here are very similar to those in the consistency test presented above. Namely, in both of the two panels, (A) and (B), we can see that the simulated rejection probabilities are dominated by the nominal size 0.05 for all of $\xi_U \in \{0.09, 0.19, 0.29, 0.39\}$ in the null region, and those are approximately the same as the nominal size 0.05 near the boundary, i.e., $\xi_U = 0.49$, of the null region. Again, these results support the size control property of the test that is established in Theorem 1.

(A) Linear Regression Model

	$n = 10^4$			$n = 10^5$			$n = 10^6$		
ξ_U	k =50	k =100	k =200	k =50	k =100	k =200	k =50	k =100	k =200
0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.79	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.99	0.06	0.06	0.07	0.05	0.05	0.06	0.05	0.04	0.05
1.19	0.19	0.28	0.42	0.19	0.25	0.39	0.17	0.24	0.39
1.39	0.39	0.57	0.70	0.39	0.57	0.78	0.38	0.56	0.80
1.59	0.59	0.70	0.69	0.59	0.80	0.91	0.59	0.82	0.96
1.79	0.69	0.73	0.59	0.75	0.88	0.88	0.77	0.93	0.96
1.99	0.75	0.68	0.51	0.85	0.89	0.83	0.87	0.96	0.94

(B) Linear IV Model

	$n = 10^4$			$n = 10^5$			$n = 10^6$		
ξ_U	k =50	k =100	k =200	k =50	k =100	k =200	k =50	k =100	k =200
0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.79	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.99	0.06	0.06	0.08	0.05	0.05	0.06	0.05	0.04	0.05
1.19	0.19	0.28	0.42	0.18	0.26	0.40	0.19	0.24	0.38
1.39	0.40	0.56	0.68	0.38	0.57	0.79	0.37	0.57	0.82
1.59	0.59	0.70	0.66	0.59	0.80	0.90	0.59	0.82	0.95
1.79	0.69	0.73	0.57	0.75	0.88	0.87	0.76	0.93	0.96
1.99	0.74	0.66	0.47	0.84	0.88	0.81	0.86	0.96	0.94

Table 2: Rejection probabilities of the test (5.3) of the consistency of $\hat{\theta}$ in (A) the linear regression model and (B) the linear IV model. The results are based on 5000 simulation draws. The significance level is 0.05.

(A) Linear Regression Model

	$n = 10^4$			$n = 10^5$			$n = 10^6$		
ξ_U	k =50	k =100	k =200	k =50	k =100	k =200	k =50	k =100	k =200
0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.39	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
0.49	0.05	0.06	0.07	0.05	0.05	0.05	0.05	0.04	0.04
0.59	0.17	0.27	0.44	0.17	0.25	0.37	0.17	0.23	0.36
0.69	0.39	0.57	0.84	0.37	0.56	0.81	0.37	0.56	0.80
0.79	0.60	0.83	0.97	0.59	0.82	0.98	0.59	0.82	0.97
0.89	0.76	0.94	0.99	0.76	0.95	1.00	0.76	0.95	1.00
0.99	0.87	0.98	0.99	0.87	0.99	1.00	0.87	0.99	1.00

(B) Linear IV Model

	$n = 10^4$			$n = 10^5$			$n = 10^6$		
ξ_U	k =50	k =100	k =200	k =50	k =100	k =200	k =50	k =100	k =200
0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.39	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
0.49	0.05	0.05	0.08	0.05	0.04	0.04	0.04	0.04	0.04
0.59	0.18	0.26	0.44	0.18	0.24	0.37	0.16	0.23	0.37
0.69	0.38	0.58	0.83	0.37	0.56	0.82	0.36	0.57	0.81
0.79	0.59	0.83	0.98	0.58	0.82	0.98	0.59	0.82	0.97
0.89	0.76	0.94	0.99	0.77	0.95	1.00	0.76	0.94	1.00
0.99	0.88	0.98	0.98	0.88	0.98	1.00	0.88	0.99	1.00

Table 3: Rejection probabilities of the test (5.3) of the asymptotic normality of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ in (A) the linear regression model and (B) the linear IV model. The results are based on 5000 simulation draws. The significance level is 0.05.

7 Application to Demand Estimation

In this section, we present an empirical application of the proposed test procedure. Recall the framework of demand estimation in differentiated products markets introduced in Example 1. The dependent variable is defined by the logarithm of the market share of a product relative to that of an outside product. In rich data sets, we often encounter zero empirical market shares. Since the logarithm of zero is undefined, empirical practitioners often use *ad hoc* procedures to deal with observations with zero market share. One common way is to simply remove observations with zero empirical market shares. Another common way is to replace zeros with a small positive value. Both of these two *ad hoc* treatments result in biased estimates in general, as demonstrated through Monte Carlo simulation studies by Gandhi, Lu, and Shi (2017). In implementing the second approach, empirical researchers often substitute infinitesimal positive values Δ for zeros, perhaps in efforts to mitigate such biases. In this paper, we show that substitution of infinitesimal positive values Δ in fact results in pathetic asymptotic behaviors of the estimator. Specifically, such an *ad hoc* estimator fails the root-n asymptotic normality, as we reject the finite second moment condition of the score. Furthermore, such an estimator is not even likely to converge in probability to a possibly biased pseudo-true target either, as we reject the finite first moment condition of the score too. These results follow because the introduction of a huge negative number (as the logarithm of an infinitesimal number) turns some of the observations with *originally non-zero* shares into outliers, as we will carefully illustrate ahead after presenting the test results.

Following preceding papers on market analysis, we use scanner data from the Dominick's Finer Foods (DFF) retail chain.[4] The unit of observation is defined by the product of UPC (universal product code), store, and week. Our analysis, as described below, follows that of Gandhi, Lu, and Shi (2017). We focus on the product category of canned tuna. Empirical

[4]We thank James M. Kilts Center, University of Chicago Booth School of Business for allowing us to use this data set. It is available at https://www.chicagobooth.edu/research/kilts/datasets/dominicks.

market shares are constructed by using quantity sales and the number of customers who visited the store in the week. Control variables include the price, UPC fixed effects, and a time trend. We instrument the possibly endogenous prices by the wholesale costs, which are calculated by inverting the gross margin.

The number of observations is approximately 10^6, similar to the sample sizes considered in our Monte Carlo simulation studies in Section 6. This feature of the data allows us to use a reasonably large number k of order statistics to enhance the power of our proposed test. Among this large number of observations, approximately 44% of the observations are recorded to have zero empirical market share. The smallest non-zero empirical market share is approximately 10^{-5}. Therefore, it is sensible to replace the zero empirical market share by an infinitesimal positive number Δ that is no larger than 10^{-5}. In our analysis, therefore, we consider the following numbers to replace zero with: $\Delta = 10^{-5}, 10^{-6}, ..., 10^{-19}, 10^{-20}$.

Table 4 summarizes the p-values of testing the finite first moment condition for the consistency. Similarly, Table 5 summarizes the p-values of testing the finite second moment condition for the root-n asymptotic normality. For the sake of transparency, we show results for various numbers of k ranging from 1000 to 5000. Before discussing these results, first note that small numbers k of order statistics in general entail short power. In view of Figure 1, we can see that k ranging from 1000 to 5000 yields very strong powers of the test. Furthermore, note also that the number $k = 5000$ corresponds to only 0.5 percent of the whole sample, so that the extreme value approximation should perform well. With these in mind, observe that the results reported in Tables 4 and 5 suggest that we start to reject the null hypothesis of a finite first and second moments when k is greater than or equal to 2000. The rejection of the finite second moment conditions (Table 5) implies that the root-n asymptotic normality of the demand estimator may perform poorly if we conduct the *ad hoc* practice of replacing the zero empirical market share by any of the infinitesimal positive values $\Delta = 10^{-5}, 10^{-6}, ..., 10^{-19}, 10^{-20}$. Furthermore, the rejection of the finite first moment conditions (Table 4) implies that such an *ad hoc* estimator

Δ	k =1000	k=2000	k =3000	k =4000	k =5000
10^{-5}	0.54	0.00	0.00	0.97	1.00
10^{-6}	1.00	0.00	0.00	0.00	0.28
10^{-7}	1.00	0.00	0.00	0.00	0.00
10^{-8}	1.00	0.00	0.00	0.00	0.00
10^{-9}	1.00	0.00	0.00	0.00	0.00
10^{-10}	1.00	0.00	0.00	0.00	0.00
10^{-11}	1.00	0.00	0.00	0.00	0.00
10^{-12}	1.00	0.00	0.00	0.00	0.00
10^{-13}	1.00	0.00	0.00	0.00	0.00
10^{-14}	1.00	0.00	0.00	0.00	0.00
10^{-15}	1.00	0.00	0.00	0.00	0.00
10^{-16}	1.00	0.00	0.00	0.00	0.00
10^{-17}	1.00	0.00	0.00	0.00	0.00
10^{-18}	1.00	0.00	0.00	0.00	0.00
10^{-19}	1.00	0.00	0.00	0.00	0.00
10^{-20}	1.00	0.00	0.00	0.00	0.00

Table 4: P-values of the test (5.3) of the finite first moment condition for consistency with the market share data from DFF for the product category of canned tuna data, where zero empirical market shares are replaced by $\Delta = 10^{-5}$, 10^{-6}, ..., 10^{-19}, 10^{-20}.

may not even converge in probability to a possibly biased pseudo-true target.

While the test rejects the null hypotheses of finite moments of $A_i^1(\theta_0)$ and $A_i^2(\theta_0)$, a natural question is why the *ad hoc* procedure of adding a small constant to the zero market share causes the heavy-tailed distributions of $A_i^1(\theta_0)$ and $A_i^2(\theta_0)$. Since the logarithm of a small constant is

Δ	k =1000	k=2000	k =3000	k =4000	k =5000
10^{-5}	0.00	0.00	0.00	0.00	0.00
10^{-6}	0.00	0.00	0.00	0.00	0.00
10^{-7}	0.01	0.00	0.00	0.00	0.00
10^{-8}	0.02	0.00	0.00	0.00	0.00
10^{-9}	0.02	0.00	0.00	0.00	0.00
10^{-10}	0.05	0.00	0.00	0.00	0.00
10^{-11}	0.06	0.00	0.00	0.00	0.00
10^{-12}	0.08	0.00	0.00	0.00	0.00
10^{-13}	0.11	0.00	0.00	0.00	0.00
10^{-14}	0.12	0.00	0.00	0.00	0.00
10^{-15}	0.14	0.00	0.00	0.00	0.00
10^{-16}	0.16	0.00	0.00	0.00	0.00
10^{-17}	0.17	0.00	0.00	0.00	0.00
10^{-18}	0.19	0.00	0.00	0.00	0.00
10^{-19}	0.21	0.00	0.00	0.00	0.00
10^{-20}	0.21	0.00	0.00	0.00	0.00

Table 5: P-values of the test (5.3) of the finite second moment condition for the root-n asymptotic normality with the market share data from DFF for the product category of canned tuna data, where zero empirical market shares are replaced by $\Delta = 10^{-5}$, 10^{-6}, ..., 10^{-19}, 10^{-20}.

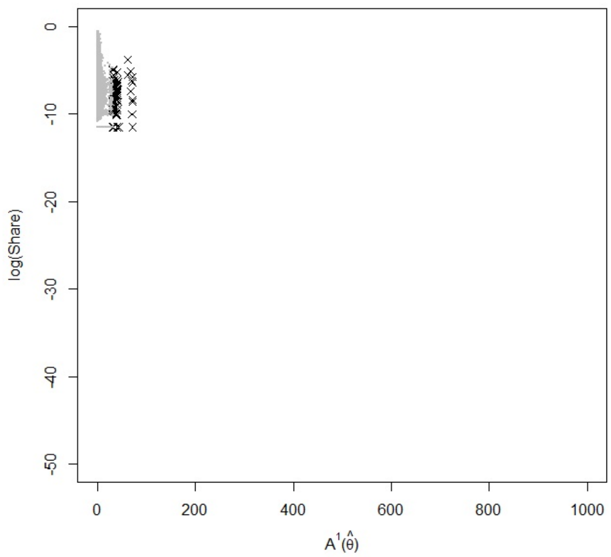
finite anyway, it appears to only produce a 44% point mass of absolutely very large yet finite constants. As such, these small numbers do not seem to contribute to heavy tails *by themselves*. To see what is going on behind our test rejecting the null hypotheses, we display eight scatter plots in Figures 2 and 3. Figure 2 displays plots of (A) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-5}$; (B) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-5}$; (C) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-10}$; and (D) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-10}$. Figure 3 displays plots of (A) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-15}$; (B) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-15}$; (C) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-20}$; and (D) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-20}$. Those observations above the top 0.0001-quantile of $A^1(\hat{\theta})$ and $A^2(\hat{\theta})$ are marked by black crosses, while all else are marked by gray dots.

In each of the panels in Figures 2 and 3, note that the observations with originally zero market share appear on the horizontal line at the vertical level of $\log(\Delta)$. As Δ becomes smaller, these lines move downward and they tend to behave as observations with an absolutely large Y value. However, these observations with originally zero shares are not necessarily outliers by themselves because as many as 44% of the observations exist on this line. Instead, many of the outliers (i.e., observations marked by the black crosses) stem from the group of observations with *originally non-zero* market shares. Furthermore, the *horizontal* distances between those marked by the black crosses and the major cluster of observations marked by gray dots widen as Δ becomes smaller, i.e., the *horizontal* spread is the smallest in Figure 2 (A)–(B) and the largest in Figure 3 (C)–(D). This pattern implies that, while most of the 44% of observations with originally zero market share are not outliers by themselves despite the isolated levels of $\log(\Delta)$, smaller values of Δ are turning some of the observations with originally non-zero market shares into outliers to larger extents. These results that the residuals of originally non-zero market shares move 'more' with Δ imply that the observations with originally zero market shares exercise more leverage in the regression estimation.

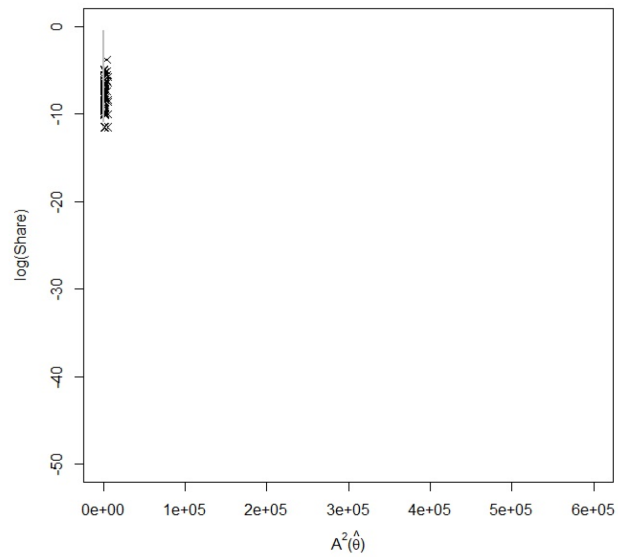
8 Summary and Discussions

Many empirical studies in economics rely on the GMM and M estimators including, but not limited to, the OLS, GLS, QMLE, and 2SLS. Furthermore, they usually rely on the consistency and the root-n asymptotic normality of these estimators when drawing scientific conclusions

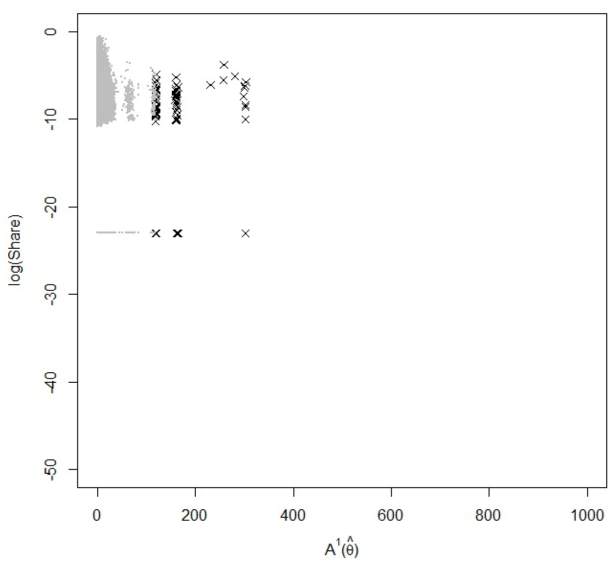
(A) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-5}$



(B) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-5}$



(C) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-10}$



(D) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-10}$

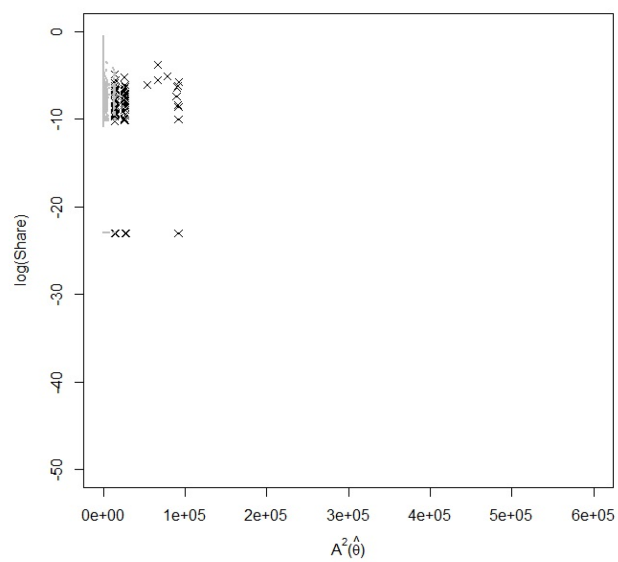


Figure 2: Scatter plots of (A) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-5}$; (B) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-5}$; (C) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-10}$; and (D) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-10}$. Observations with the top 0.0001-quantile of $A^1(\hat{\theta})$ and $A^2(\hat{\theta})$ are marked by black crosses.

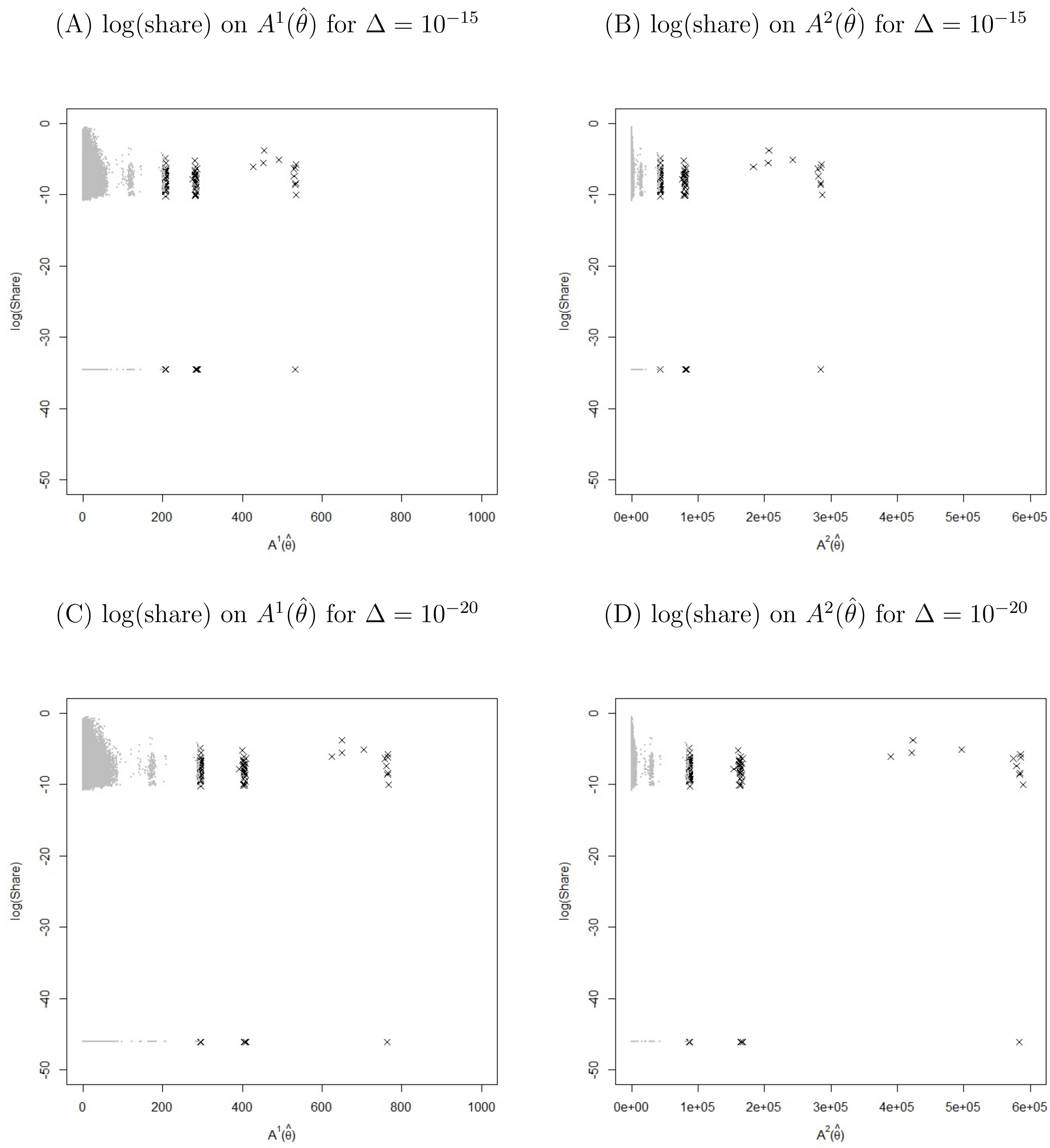


Figure 3: Scatter plots of (A) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-15}$; (B) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-15}$; (C) log(share) on $A^1(\hat{\theta})$ for $\Delta = 10^{-20}$; and (D) log(share) on $A^2(\hat{\theta})$ for $\Delta = 10^{-20}$. Observations with the top 0.0001-quantile of $A^1(\hat{\theta})$ and $A^2(\hat{\theta})$ are marked by black crosses.

via statistical inference. Although the conditions for the consistency and the root-n asymptotic normality are usually taken for granted as such, they may not be always plausibly satisfied. In this light, this paper proposes a method of testing the hypothesis of finite first and second moments of scores, which serve as key conditions of the consistency and the root-n asymptotic normality, respectively.

There are two desired properties of our proposed test in practice. First, unlike other approaches in extreme value theory that require a sequence of tuning parameter values that change as the sample size grows, our test is valid for any predetermined fixed number k of order statistics to be used to construct the test. This is a useful property in practice because it relieves researchers from worrying about a 'valid' data driven choice of tuning parameters for the purpose of size control. Second, our test has a size control property over the set of data generating processes for which the finite moment condition holds. Monte Carlo simulation studies indeed support this theoretical property for two of the most commonly used econometric frameworks, namely the linear regression model and the linear IV model.

A failure of the consistency and the root-n asymptotic normality may be caused by the following two cases among others. First, some dependent variables (e.g., wealth, infant birth weight, murder rate, city size, stock returns) are reported to exhibit heavy-tailed distributions, and they can induce infinite first and second moments of the score of an estimator. Second, when a dependent variable is the logarithm of a variable, practitioners sometimes employ an *ad hoc* procedure of replacing zeros by infinitesimal values. This practice can lead to heavy-tailed distribution of the residuals for observations with originally non-zero market shares. In our empirical application, we highlighted the latter case. Using scanner data from the Dominick's Finer Foods (DFF) retail chain, we reject the consistency and the root-n asymptotic normality for demand estimators based on such an *ad hoc* practice.

Finally, we conclude this paper by remarking that the test can be used to enhance the quality and credibility of past and future empirical studies. On one hand, if our test supports the finite moment conditions for consistency and the root-n asymptotic normality for a selected empirical work, then the test result reinforces the credibility of scientific conclusions reported by that work. On the other hand, if our test fails to support the finite moment conditions for a selected empirical work, then a researcher may want to consider one of the alternative robust approaches for more credible empirical research.

References

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007): "Econometric tools for analyzing market outcomes," *Handbook of Econometrics*, 6, 4171–4276.

Al-Sadoon, M. M. (2017): "A unifying theory of tests of rank," *Journal of Econometrics*, 199, 49–62.

Andrews, D. W. K. and W. Ploberger (1994): "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

Bahadur, R. R. and L. J. Savage (1956): "The nonexistence of certain statistical procedures in nonparametric problems," *The Annals of Mathematical Statistics*, 27, 1115–1122.

Barro, R. J. and T. Jin (2011): "On the Size Distribution of Macroeconomic Disasters," *Econometrica*, 79, 1567–1589.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, 841–890.

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *RAND Journal of Economics*, 242–262.

Camba-Méndez, G. and G. Kapetanios (2009): "Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling," *Econometric Reviews*, 28, 581–611.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2014): "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory*, 30, 176–200.

Cragg, J. G. and S. G. Donald (1996): "On the asymptotic properties of LDU-based tests of the rank of a matrix," *Journal of the American Statistical Association*, 91, 1301–1309.

——— (1997): "Inferring the rank of a matrix," *Journal of econometrics*, 76, 223–250.

Davis, R. and S. Resnick (1985): "More limit theory for the sample correlation function of moving averages," *Stochastic Processes and their Applications*, 20, 257–279.

——— (1986): "Limit theory for the sample covariance and correlation functions of moving averages," *Annals of Statistics*, 533–558.

Davis, R. A., K. Knight, and J. Liu (1992): "M-estimation for autoregressions with infinite variance," *Stochastic Processes and Their Applications*, 40, 145–180.

De Haan, L. and A. Ferreira (2006): *Extreme Value Theory: An Introduction*, New York: Springer Science and Business Media.

Edgeworth, F. Y. (1887): "On observations relating to several quantities," *Hermathena*, 6, 279–285.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly optimal tests when a nuisance parameter is present under the null hypothesis," *Econometrica*, 83, 771–811.

Gabaix, X. (2016): "Power Laws in Economics: An Introduction," *Journal of Economic Literature*, 30, 185–206.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley (2003): "A theory of power-law distributions in financial market fluctuations," *Nature*, 423, 267–270.

Gandhi, A., Z. Lu, and X. Shi (2017): "Estimating demand for differentiated products with zeroes in market share data," *Microsoft Research, SUFE, and UW-Madison.*

Gill, L. and A. Lewbel (1992): "Testing the rank and definiteness of estimated matrices with applications to factor, state-space and ARMA models," *Journal of the American Statistical Association*, 87, 766–776.

Hill, J. B. and A. Prokhorov (2016): "GEL estimation for heavy-tailed GARCH models with robust empirical likelihood inference," *Journal of Econometrics*, 190, 18–45.

Huber, P. J. (1992): "Robust estimation of a location parameter," in *Breakthroughs in statistics*, Springer, 492–518.

Inoue, A. and B. Rossi (2011): "Testing for weak identification in possibly nonlinear models," *Journal of Econometrics*, 161, 246–261.

Kiefer, N. and T. J. Vogelsang (2005): "A new asymptotic theory for heteroskedasticity-autocorrelation robust tests," *Econometric Theory*, 21, 1130–1164.

Kleibergen, F. and R. Paap (2006): "Generalized reduced rank tests using the singular value decomposition," *Journal of econometrics*, 133, 97–126.

Koenker, R. and G. Bassett Jr (1978): "Regression quantiles," *Econometrica*, 33–50.

Lee, D. L., J. McCrary, M. J. Moreira, and J. Porter (2020): "Valid t-ratio Inference for IV," *arXiv preprint arXiv:2010.05058*.

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

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

Müller, U. K. (2020): "A more robust t-test," *Working Paper*.

Müller, U. K. and Y. Wang (2017): "Fixed-k asymptotic inference about tail properties," *Journal of the American Statistical Association*, 112, 1334–1343.

Newey, K. and D. McFadden (1994): "Large sample estimation and hypothesis," *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Piketty, T. and E. Saez (2003): "Income inquality in the United States, 1913-1998," *The Quarterly Journal of Economics*, 118, 1–41.

Resnick, S. (2007): *Heavy-tail phenomena: probabilistic and statistical modeling*, Springer Science & Business Media.

Robin, J.-M. and R. J. Smith (2000): "Tests of rank," *Econometric Theory*, 16, 151–175.

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

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse (2011): "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*, 101, 2205–25.

Sanderson, E. and F. Windmeijer (2016): "A weak instrument F-test in linear IV models with multiple endogenous variables," *Journal of Econometrics*, 190, 212–221.

Shao, Q.-M., H. Yu, and J. Yu (2001): "Do stock returns follow a finite variance distribution?" *Annals of Economics and Finance*, 2, 467.

Stock, J. H. and M. Yogo (2005): "Testing for weak instruments in linear IV regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, 80–108.

Wright, J. H. (2003): "Detecting lack of identification in GMM," *Econometric Theory*, 19, 322–330.

Supplementary Appendix of

"Diagnostic Testing of Finite Moment Conditions for the Consistency and Root-N Asymptotic Normality of the GMM and M Estimators"

Yuya Sasaki* and Yulong Wang†

Abstract

This supplementary material consists of two appendix sections. Appendix A contains proofs of the main results, namely Lemma 1, Lemma 2, and Theorem 1. Appendix B contains additional computational details to implement the test.

*Associate professor of economics, Vanderbilt University. Email: yuya.sasaki@vanderbilt.edu

†Assistant professor of economics, Syracuse University. Email: ywang402@maxwell.syr.edu.

Appendix

A Proofs

A.1 Proof of Lemma 1

Proof. First consider $\xi < 1$. The result follows readily from Karamata's theorem (e.g., Resnick, 2007, Theorem 2.1 with $\rho = -1/\xi < -1$). Specifically, we have that

$$1 - F_A(a) = a^{-1/\xi}\mathcal{L}(a),$$

where $\mathcal{L}(a) \to 1$ as $a \to \infty$ (e.g., Resnick, 2007, Ch.2). Using integration by parts, we have that

$$\begin{aligned}\mathbb{E}\left[A\right] &= -\int_0^\infty a d\left(1 - F_A\left(a\right)\right) \\ &= -\lim_{a\to\infty} a\left(1 - F_A\left(a\right)\right) + \int_0^\infty \left(1 - F_A\left(a\right)\right) da.\end{aligned}$$

Both of the two terms on the right-hand side above are finite as implied by Karamata's theorem.

Next, we establish that $\xi > 1$ implies $\mathbb{E}\left[A\right] = \infty$. To this end, let $Q_A\left(s\right)$ denote the quantile function of A. The regular variation condition implies that $Q_A\left(1 - 1/n\right) = O\left(n^{\xi}\right)$ as $n \to \infty$. Then

$$\begin{aligned}\mathbb{E}\left[A\right] &= \int_0^1 Q_A\left(s\right) ds \\ &\geq n^{-1} Q_A\left(1 - 1/n\right) \\ &= O\left(n^{\xi-1}\right) \to \infty \text{ as } n \to \infty.\end{aligned}$$

Now, it remains to consider the boundary case with $\xi = 1$. Since $\mathcal{L}(\cdot)$ is bounded below (say, by $c > 0$), we have

$$\begin{aligned}
\mathbb{E}[A] &= -\int_0^\infty ad\left(1 - F_A(a)\right) \\
&= -\lim_{a\to\infty} a\left(1 - F_A(a)\right) + \int_0^\infty \left(1 - F_A(a)\right) da \\
&= -\lim_{a\to\infty} \mathcal{L}(a) + \int_0^\infty a^{-1/\xi}\mathcal{L}(a)\, da \\
&\geq -1 + c\int_0^\infty a^{-1} da = \infty.
\end{aligned}$$

This completes the proof. □

A.2 Proof of Lemma 2

Proof. First, by extreme value theory, Condition 1.(i) (D_i is i.i.d.) and Condition 1.(ii) ($F_{A^r(\theta_0)} \in \mathcal{D}(G_\xi)$) imply

$$\frac{\mathbf{A}^r(\theta_0) - b_n}{a_n} \xrightarrow{d} \mathbf{V}, \tag{A.1}$$

where $\mathbf{V}$ is jointly extreme value distributed with tail index ξ. By Corollary 1.2.4 and Remark 1.2.7 in De Haan and Ferreira (2006), the constants a_n and b_n can be chosen as $a_n = Q_{A^r(\theta_0)}(1 - 1/n) = O(n^\xi)$ and $b_n = 0$. By construction, these constants satisfy that $1 - F_{A^r(\theta_0)}(a_n y + b_n) = O\left(n^{-1}\right)$ for every $y > 0$.

Now, let $I = (I_1, \ldots, I_k) \in \{1, \ldots, n\}^k$ be the k random indices such that $A^r_{(j)}(\theta_0) = A^r_{I_j}(\theta_0)$, $j = 1, \ldots, k$, and let $\hat{I}$ be the corresponding indices such that $A^r_{(j)}\left(\hat{\theta}\right) = A^r_{\hat{I}_j}\left(\hat{\theta}\right)$. Then, the convergence of $\mathbf{A}^r\left(\hat{\theta}\right)$ follows from (A.1) once we establish $|A^r_{\hat{I}_j}(\hat{\theta}) - A^r_{I_j}(\theta_0)| = o_p(a_n)$ for $j = 1, \ldots, k$. We present the case of $k = 1$, but the argument for a general k is similar. Denote $\varepsilon_i \equiv A^r_i(\hat{\theta}) - A^r_i(\theta_0)$.

Finally, Condition 1.(iii) yields that

$$\begin{aligned}
\sup_i |\varepsilon_i| &= \sup_i \left| A^r_i(\hat{\theta}) - A^r_i(\theta_0) \right| \\
&\leq \sup_i \sup_{\theta \in B_{\eta_n}(\theta_0)} \left|\left| \frac{\partial A^r_i(\theta)}{\partial \theta} \right|\right| \left|\left| \hat{\theta} - \theta_0 \right|\right| \\
&= o_p(a_n).
\end{aligned}$$

Given this result, we have that, on one hand, $A^r_{\hat{I}}(\hat{\theta}) = \max_i\{A^r_i(\theta_0) + \varepsilon_i\} \leq A^r_I(\theta_0) + \sup_i |\varepsilon_i| = A^r_I(\theta_0) + o_p(a_n)$; and, on the other hand, $A^r_{\hat{I}}(\hat{\theta}) = \max_i\{A^r_i(\theta_0) + \varepsilon_i\} \geq \max_i\{A^r_i(\theta_0) + \min_i\{\varepsilon_i\}\} \geq A^r_I(\theta_0) + \min_i\{\varepsilon_i\} \geq A^r_I(\theta_0) - \sup_i |\varepsilon_i| = A^r_I(\theta_0) - o_p(a_n)$. Therefore, $|A^r_{\hat{I}_j}(\hat{\theta}) - A^r_{I_j}(\theta_0)| \leq o_p(a_n)$ holds. □

A.3 Proof of Theorem 1

Proof. By Lemma 1 and the continuous mapping theorem, we have $\mathbf{A}^r_*\left(\hat{\theta}\right) \xrightarrow{d} \mathbf{V}_*$. Since the density $f_{\mathbf{V}_*}$ is continuous, $\mathbb{E}_\xi\left[\varphi\left(\mathbf{V}_*\right)\right]$ as a function of ξ and cv is also continuous in both arguments for any given $\Lambda(\cdot)$. Therefore, we can choose a large enough cv so that $\sup_{\xi\in(0,1)} \mathbb{E}_\xi\left[\varphi\left(\mathbf{V}_*\right)\right] \leq \alpha$. □

Remark 1. *Since Λ in the last part of the above proof can be arbitrary in theory, we provide an empirical guide for determining a nearly optimal Λ in the following section.*

B Computational Details

This section provides computational details about constructing the test (5.3), which is based on the limit observation $\mathbf{V}_*$. The density of $\mathbf{V}_*$ is given by (5.2), which is computed by Gaussian Quadrature. To construct the test (5.3), we specify the weight W to be the uniform distribution for simplicity of exposition. The weight W reflects the importance attached by the econometrician to different alternatives, which can be easily changed. Then, it remains to determine a suitable candidate for the weight Λ. We do this by the generic algorithm provided by Elliott, Müller, and Watson (2015).

The idea is as follows. For expositional ease, we can always subsume cv into Λ, which now does not necessarily integrate to one. First, we can discretize Ξ into a grid Ξ_a and determine Λ accordingly as the point masses. Then we can simulate N random draws of $\mathbf{V}_*$ from $\xi \in \Xi_a$ and estimate $\mathbb{P}_\xi(\varphi_\Lambda(\mathbf{V}_*) = 1)$ by sample fractions, where the subscripts ξ and Λ respectively emphasize that the rejection probability depends on the ξ that generates the data and the test depends on Λ. By iteratively increasing or decreasing the point masses as a function of whether the estimated $\mathbb{P}_\xi(\varphi_\Lambda(\mathbf{V}_*) = 1)$ is larger or smaller than the nominal level, we can always find a candidate Λ^*. Note that such Λ^* always exists since we allow $\mathbb{P}_\xi(\varphi_{\Lambda^*}(\mathbf{V}_*) = 1) < \alpha$ for some ξ. The continuity of $f_{\mathbf{V}_*}$ entails that $\mathbb{P}_\xi(\varphi_\Lambda(\mathbf{V}_*) = 1)$ as a function of ξ is also continuous. Therefore, the size control over $\xi \in (0, 1)$ is guaranteed as we consider a fine enough grid Ξ_a.

In practice, we can determine the point masses by the following concrete steps. It turns out that Λ puts all mass on the point $\xi = 1$.

Algorithm:

1. Simulate $N = 10{,}000$ i.i.d. random draws from some proposal density with ξ drawn uniformly from Ξ_a, which is an equally spaced grid on $[0.01, 1]$ with 50 points.

2. Start with $\Lambda_{(0)} = \{1/50, 1/50, \ldots, 1/50\}^\intercal$. Calculate the (estimated) rejection probabilities $P_j = \mathbb{P}_{\xi_j}(\varphi_{\Lambda_{(0)}}(\mathbf{V}_*) = 1)$ for every $\xi_j \in \Xi_a$ using importance sampling. Denote them by $P = (P_1, ..., P_{50})^\intercal$.

3. Update Λ by setting $\Lambda_{(s+1)} = \Lambda_{(s)} + \eta_\Lambda(P - 0.05)$ with some step-length constant $\eta_\Lambda > 0$, so that the j-th point mass in Λ is increased/decreased if the rejection probability for ξ_j is larger/smaller than the nominal level.

4. Keep the integration for 500 times. Then, the resulting $\Lambda_{(500)}$ is a valid candidate. Then normalize $\Lambda_{(500)}$ to obtain cv.

5. Numerically check if $\varphi_{\Lambda_{(500)}}$ indeed controls the size uniformly by simulating the rejection probabilities over a much finer grid on Ξ. If not, go back to step 2 with a finer Ξ_a.

References

De Haan, L. and A. Ferreira (2006): *Extreme Value Theory: An Introduction*, New York: Springer Science and Business Media.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly optimal tests when a nuisance parameter is present under the null hypothesis," *Econometrica*, 83, 771–811.

Resnick, S. (2007): *Heavy-tail phenomena: probabilistic and statistical modeling*, Springer Science & Business Media.