Journal of Econometrics 167 (2012) 1–15

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Optimal inference for instrumental variables regression with non-Gaussian errors[☆]

Matias D. Cattaneo[a], Richard K. Crump[b], Michael Jansson[c,d,*]

[a] Department of Economics, University of Michigan, United States
[b] Federal Reserve Bank of New York, United States
[c] Department of Economics, UC Berkeley, United States
[d] CREATES, University of Aarhus, Denmark

ARTICLE INFO

Article history:
Received 25 February 2008
Received in revised form
21 April 2011
Accepted 25 April 2011
Available online 31 October 2011

ABSTRACT

This paper is concerned with inference on the coefficient on the endogenous regressor in a linear instrumental variables model with a single endogenous regressor, nonrandom exogenous regressors and instruments, and *i.i.d.* errors whose distribution is unknown. It is shown that under mild smoothness conditions on the error distribution it is possible to develop tests which are "nearly" efficient in the sense of Andrews et al. (2006) when identification is weak and consistent and asymptotically optimal when identification is strong. In addition, an estimator is presented which can be used in the usual way to construct valid (indeed, optimal) confidence intervals when identification is strong. The estimator is of the two stage least squares variety and is asymptotically efficient under strong identification whether or not the errors are normal.

1. Introduction

This paper is concerned with inference on the coefficient on the endogenous regressor in a linear instrumental variables (IVs) model with a single endogenous regressor, nonrandom exogenous regressors and IVs, and *i.i.d.* errors. Models of this type have been studied intensively in recent years, with particular attention being devoted to the case where the IVs are weak (in the terminology of Staiger and Stock (1997)).[1] Analyzing such a model in which the *i.i.d.* errors are furthermore assumed to be Gaussian, Andrews et al. (2006, henceforth AMS) find that the conditional likelihood ratio test proposed by Moreira (2003) is "nearly" efficient when identification is weak and asymptotically efficient when identification is strong.

The purpose of the present paper is to explore the consequences of relaxing the assumption of normality on the part of the *i.i.d.* errors in a model which is otherwise identical to the model studied by AMS (and others). Recent work by Andrews and Marmer (2007) and Andrews and Soares (2007) shows that departures from normality can be exploited for power purposes when the errors satisfy a certain symmetry condition. Although these papers do not establish optimality results on the part of the rank-based testing procedures proposed therein, the findings of the papers imply in particular that for certain classes of error distributions the conditional likelihood test ceases to be (nearly) optimal once the assumption of normality is relaxed. This paper addresses the issue of optimality and shows that under mild smoothness conditions on the (otherwise unknown) error distribution it is possible to develop tests which are (nearly) optimal whether or not the errors are Gaussian.

The asymptotic optimality theory developed herein treats the distribution of the *i.i.d.* errors as an unknown nuisance parameter and is therefore of the semiparametric variety. In fact, under the assumption that the model contains an intercept (an assumption which we maintain throughout), we establish adaptation results, namely that one can construct procedures which perform asymptotically as well as procedures which optimally utilize knowledge of the error distribution. This adaptation result bears more than a superficial resemblance to Bickel's (1982) celebrated result on adaptive estimation of the slope coefficients in a regression model. Specifically, it turns out that the problem of conducting inference in an IV model with an unknown error distribution can

[☆] The authors thank Don Andrews, Bryan Graham, Jim Powell, Tom Rothenberg, Paul Ruud, Hal White, two anonymous referees, and seminar participants at Aarhus, Berkeley, Brown, Columbia, Michigan, NYU, Princeton, Rochester, Stanford, Yale, UCSD, UTDT, and UdeSA for comments. The first author gratefully acknowledges financial support from the National Science Foundation (SES 0921505). The third author gratefully acknowledges financial support from the National Science Foundation (SES 0920953) and the research support of *CREATES* (funded by the Danish National Research Foundation).

[*] Corresponding author. Tel.: +1 510 642 4639.
E-mail address: mjansson@econ.berkeley.edu (M. Jansson).

[1] For reviews of the weak IV literature, see e.g. Dufour (2003) and Andrews and Stock (2007).

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.04.004

be decomposed into two separate problems, each of which is well understood, in isolation, from the works of Bickel (1982) and AMS, respectively. The first of these problems concerns efficient estimation of the slope coefficients in the reduced form of the IV model. That problem is a bivariate version of the problem addressed by Bickel (1982) and can be solved in essentially the same way. Because efficient estimators of the slope coefficients turn out to be asymptotically sufficient statistics for the relevant parameters of the IV model, the problem of conducting optimal inference can be reduced to the problem of optimally extracting information from the efficient estimators of the reduced form regression coefficients. The mathematical structure of that problem turns out to be the same whether or not the errors are Gaussian, implying that we can utilize the results of AMS to construct test statistics which combine the efficient estimators of the reduced form regression coefficients in an optimal way.

Our construction of feasible inference procedures proceeds in several steps, culminating with a procedure which is nearly efficient when identification is weak and consistent and asymptotically optimal when identification is strong. The resulting procedure is of the conditional likelihood ratio variety, but being optimal (or nearly so, depending on the strength of identification) it is of necessity different from Moreira's (2003) procedure. Analogously to Moreira's (2003) procedure, a potential drawback of our procedure is that although it enjoys optimality properties when identification is strong, it is somewhat tedious to invert it in order to obtain confidence intervals in strongly identified models. To address this issue, we present an estimator and an accompanying standard error formula which can be used in the usual way to construct valid (indeed, optimal) confidence intervals when identification is strong. The estimator, which would appear to be new, is of the two stage least squares (2SLS) variety and is asymptotically efficient under strong identification whether or not the errors are normal.

The paper proceeds as follows. Section 2 presents the model and the assumptions under which the asymptotic analysis will proceed. Section 3 is concerned with asymptotic inference under the assumptions that the error distribution is known and identification is weak. The counterfactual assumption that the error distribution is known is dispensed with in Section 4, where it is also shown how strong identification can be accommodated. Section 5 presents some simulation results, while mathematical derivations have been relegated to an Appendix.

2. The model

We consider a model given by

$$y_{1i} = \Gamma_1' x_i + \beta y_{2i} + u_i,$$

$$y_{2i} = \gamma_2' x_i + \pi' z_i + v_{2i} \quad (i = 1, \ldots, n), \tag{1}$$

where $y_{1i}, y_{2i} \in \mathbb{R}$, $x_i \in \mathbb{R}^p$, and $z_i \in \mathbb{R}^q$ are observed variables; $u_i, v_{2i} \in \mathbb{R}$ are unobserved errors; and $\beta \in \mathbb{R}$, $\pi \in \mathbb{R}^q$, and Γ_1, $\gamma_2 \in \mathbb{R}^p$ are parameters. The exogenous variables x_i and z_i are nonrandom and the first element of x_i is assumed to be equal to unity. The errors (u_i, v_{2i}) are *i.i.d.* from a continuous distribution with zero mean and finite variance.

It turns out to be convenient to work with the reduced form of the model. The reduced form is given by the pair of equations

$$y_{1i} = \gamma_1' x_i + \beta \pi' z_i + v_{1i},$$

$$y_{2i} = \gamma_2' x_i + \pi' z_i + v_{2i} \qquad (i = 1, \ldots, n), \tag{2}$$

where $\gamma_1 = \Gamma_1 + \gamma_2 \beta$ and $v_{1i} = v_{2i}\beta + u_i$. The parameters of the reduced form are $\beta, \pi, \gamma = \left(\gamma_1', \gamma_2'\right)'$, and f, the Lebesgue density of $v_i = (v_{1i}, v_{2i})'$. The analysis of the reduced form is facilitated by the fact that it can be embedded in the model

$$y_{1i} = \gamma_1' x_i + \delta_1' z_i + v_{1i},$$

$$y_{2i} = \gamma_2' x_i + \delta_2' z_i + v_{2i} \quad (i = 1, \ldots, n), \tag{3}$$

where $\delta_1, \delta_2 \in \mathbb{R}^q$ and the other parameters are as in (2).[2] Indeed, the main results of this paper can and will be derived as relatively simple consequences of results concerning the bivariate regression model (3), which itself can be analyzed by means of fairly standard tools.

Our goal is to develop powerful tests of

$$H_0 \colon \beta = \beta_0 \quad \text{vs.}\; H_1 \colon \beta \neq \beta_0,$$

treating π, γ, and f as unknown nuisance parameters.[3] Replacing y_{1i} by $y_{1i} - \beta_0 y_{2i}$ if necessary, we assume without loss of generality that $\beta_0 = 0$.

The analysis proceeds under the following assumptions.[4]

Assumption 1. (a) $Q_{zz,n} = n^{-1} \sum_{i=1}^n z_i z_i' \to Q_{zz} > 0$ and $\max_{1 \le i \le n} \|z_i\| / \sqrt{n} \to 0$. (b) $Q_{xx,n} = n^{-1} \sum_{i=1}^n x_i x_i' \to Q_{xx} > 0$ and $\max_{1 \le i \le n} \|x_i\| / \sqrt{n} \to 0$.

Assumption 2. The density f admits a function $\dot f$ such that

(a) for almost every $v \in \mathbb{R}^2$, f is differentiable at v, with total derivative $\dot f$;

(b) for every $v, \theta \in \mathbb{R}^2$, $f(v + \theta) - f(v) = \theta' \int_0^1 \dot f(v + \theta t)\, dt$;

(c) $\int_{\mathbb{R}^2} \|\ell(v)\|^2 f(v)\, dv < \infty$, where $\ell(v) = -1[f(v) > 0]\, \dot f(v)/f(v)$.

Assumption 3. $Q_{zx,n} = n^{-1} \sum_{i=1}^n z_i x_i' \to 0$.

Remarks. (i) Assumption 1 is a fairly standard assumption concerning the exogenous variables. As in Bickel (1982), the assumption that the exogenous variables $\left(x_i', z_i'\right)'$ are nonrandom can be relaxed, and the main results of this paper will remain valid, provided the errors $\{v_i\}$ are assumed to be independent of $\left\{\left(x_i', z_i'\right)'\right\}$.

(ii) The assumption that second moments of the errors exist serves three purposes. First, it implies that the Fisher information matrix $\mathcal{I}$ defined in (4) is nonsingular. Second, it implies that the $\sqrt{n}$-consistency requirements of Assumptions 6–8 are met by OLS estimators. Finally, it is required for the validity of the statements concerning procedures based on the Gaussian (quasi-)likelihood that are made throughout the paper. As in Bickel (1982), the main results of this paper are valid even without moment assumptions provided it is assumed that $\mathcal{I} > 0$.

(iii) Assumption 2 is a relatively mild smoothness condition on the error density. Parts (a) and (b) of Assumption 2 hold if, but do not require that, f is continuously differentiable. In particular, Assumption 2 accommodates mild departures from continuous differentiability, such as that which occurs when the elements of v_i, or some rotation thereof, are independent and double exponentially distributed.

(iv) If Assumption 1 holds and $Q_{zx} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n z_i x_i'$ exists, Assumption 3 is a normalization in the sense that it entails no loss of generality. Specifically, replacing z_i by $z_i^* = z_i - Q_{zx} Q_{xx}^{-1} x_i$ has no effect on the value of (β, π) and guarantees validity of Assumption 3. Our main results depend on $\left\{\left(x_i', z_i'\right)'\right\}$ only through $\left\{z_i^*\right\}$, so Assumption 3 is convenient insofar as it enables us to simplify the notation by eliminating the distinction between $\{z_i\}$ and $\left\{z_i^*\right\}$.

[2] The model (3) reduces to (2) when $\delta = (\delta_1, \delta_2)' = \left(\beta\pi', \pi'\right)'$.

[3] Testing problems of this type are of interest partly because the duality between hypothesis testing and interval estimation implies that confidence intervals for β can be obtained by test inversion.

[4] In Assumption 1 and elsewhere in the paper, $\|\cdot\|$ is the Euclidean norm and limits are taken as $n \to \infty$, except where otherwise noted.

(v) Throughout this paper, the endogenous regressor y_{2i} is assumed to be scalar. Most of our distributional results should generalize straightforwardly to models with multiple endogenous regressors, as should the optimality results reported in Section 4.4. On the other hand, analogues of the near optimality results (established by AMS for Moreira (2003)-type inference procedures in models with weak instruments and a scalar endogenous regressor) that underlie some of the efficiency claims made in other sections of the paper do not seem to be available for models with multiple endogenous regressors.

An immediate implication of Assumption 1(a) and 2 is that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \ell\left(v_i\right) \otimes z_i \rightarrow_d \mathcal{N}\left(0, \mathcal{I} \otimes Q_{zz}\right),$$

where

$$\mathcal{I} = \int_{\mathbb{R}^2} \ell\left(v\right) \ell\left(v\right)' f\left(v\right) dv \tag{4}$$

is the Fisher information for the location family generated by f. As shown in the Appendix, Assumption 2 furthermore enables nonparametric estimation of ℓ and implies that the model (3) is differentiable in quadratic mean at any (γ, δ). In other words, the roles played by parts (a) and (b) of Assumption 2 are analogous to those played by the assumption of absolute continuity routinely invoked in regression models with scalar errors. In fact, the natural scalar counterpart of Assumption 2(b) is the assumption of absolute continuity.

As mentioned in remark (iv), Assumption 3 is a normalization which greatly simplifies the derivation and statements of asymptotic results. Specifically, because the limit of $Q_{zx,n}$ is a zero matrix under Assumption 3, the parameters (β, π) and γ are orthogonal (in the sense of Cox and Reid (1987)). This fact, which is an immediate consequence of the fact that $\delta = (\delta_1, \delta_2)'$ and γ are orthogonal in (3), implies that the analysis can proceed under the "as if" assumption that γ is known. Similarly, the fact that $n^{-1}\sum_{i=1}^{n} z_i \to 0$ under Assumption 3 (because the first element of x_i equals unity) implies that the analysis can proceed under the "as if" assumption that f is known. This is so because δ in (3) can be estimated adaptively, the latter fact essentially following from Bickel's (1982) result on adaptive estimation of slope coefficients in a regression model.

In other words, Assumption 3 implies that π is the only nuisance parameter which matters asymptotically. Concerning π, particular attention will be devoted to the weakly identified case where π is "close" to zero in the sense of the following assumption.

Assumption 4W. $\pi = c/\sqrt{n}$ for some constant $c \in \mathbb{R}^q$ and β is a constant.

Under the local-to-zero parameterization of π specified by Assumption 4W, contiguous alternatives to H_0 are of the form $\beta = \beta_0 + O(1)$. Accordingly, β is modeled as a constant in the weakly identified case. Although our main emphasis is on the weakly identified case, we shall on occasion employ one of the following (strong identification) assumptions.

Assumption 4SC. π is a nonzero constant and $\beta = b/\sqrt{n}$ for some constant $b \in \mathbb{R}$.

Assumption 4SF. π is a nonzero constant and β is a constant.

When π is a nonzero constant, identification is strong and contiguous alternatives to H_0 are of the form $\beta = \beta_0 + O\left(1/\sqrt{n}\right)$. Assumption 4SC covers that case and is appropriate when studying local asymptotic power properties under strong identification. In contrast, Assumption 4SF assumes strong identification and furthermore holds β fixed. This combination of strong identification and fixed alternatives is appropriate when studying the consistency properties of various tests. Moreover, Assumption 4SF is useful when studying the properties of point estimators of β under strong identification.

Assumptions 4SC, 4SF and 4W are nonnested, but it seems natural to study them in the order indicated above. This is so because the assumptions impose decreasingly strong upper bounds on the magnitude of the parameters δ_1 and δ_2 of (3). Specifically, Assumption 4W implies that $\delta_1 = O\left(1/\sqrt{n}\right)$ and $\delta_2 = O\left(1/\sqrt{n}\right)$. Relative to Assumptions 4SC and 4W removes the requirement $\delta_2 = O\left(1/\sqrt{n}\right)$ and Assumption 4SF furthermore relaxes the requirement $\delta_1 = O\left(1/\sqrt{n}\right)$. In this paper, these differences are important because the feasible inference procedures constructed in Section 4 employ one-step estimators of δ. As usual, one-step estimators utilize initial estimators that are required to be $\sqrt{n}$-consistent. Under Assumption 4W, this requirement is met by the zero vector, while Assumptions 4SC and 4SF imply that nondegenerate initial estimators of δ_2 and (δ_1, δ_2), respectively, are required in order to guarantee that one-step estimators of δ are well behaved. Accordingly, the three constructions presented in Section 4 differ in terms of (and only in terms of) the nature of the initial estimators of δ being employed.

3. The limiting experiment when identification is weak

This section is concerned with asymptotic inference under the assumptions that (i) the nuisance parameters γ and f are known and (ii) identification is weak. As mentioned in the previous section, Assumption 3 ensures that (i) can be dispensed with. Precise statements to that effect will be provided in the next section, where it is also shown how departures from (ii) can be accommodated.

When f is Gaussian and the reduced form variance $\Omega = \int_{\mathbb{R}^2} vv'f\left(v\right) dv$ is known, the problem of testing $\beta = \beta_0 vs. \beta \neq \beta_0$ is nonstandard, but amenable to finite sample analysis using the theory of curved exponential families (e.g., Moreira (2003) and AMS). This feature is lost, in general, when f is not Gaussian. On the other hand, the testing problem remains amenable to asymptotic analysis using the limits of experiments approach even when f is non-Gaussian.[5] In fact, it turns out that the family of limiting experiments associated with non-Gaussian error distributions coincides with the family of limiting experiments for the Gaussian case.

In the Gaussian case, the limiting experiment is that of a single observation from the $\mathcal{N}\left[\mu\left(\beta, c\right), \Omega \otimes Q_{zz}^{-1}\right]$ distribution, where

$$\mu\left(\beta, c\right) = \left(\beta, 1\right)' \otimes c.$$

Equivalently, because $\Omega = \mathcal{I}^{-1}$ when f is Gaussian, the limiting experiment in the Gaussian case is that of a single observation from the $\mathcal{N}\left[\mu\left(\beta, c\right), \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right]$ distribution. As it turns out, the latter characterization generalizes readily to non-Gaussian error distributions.

To give a precise statement, we proceed in the spirit of van der Vaart (1998, Section 7.6). Define the log likelihood ratio function

$$L_n\left(\beta, c\right) = \sum_{i=1}^{n} \log f\left(y_{1i} - \gamma_1' x_i - \beta c' z_i/\sqrt{n}, y_{2i} - \gamma_2' x_i - c' z_i/\sqrt{n}\right) - \sum_{i=1}^{n} \log f\left(y_{1i} - \gamma_1' x_i, y_{2i} - \gamma_2' x_i\right)$$

[5] For an exposition of the elements of the theory of limits of experiments employed in this paper, see e.g. van der Vaart (1998).

and let "$o_{p_0}(1)$" and "$\to_{d_0}$" be shorthand for "$o_p(1)$ under the distributions associated with $(\beta, \pi) = (0, 0)$" and "$\to_d$ under the distributions associated with $(\beta, \pi) = (0, 0)$", respectively.

Theorem 1. *If Assumption 1 (a) and 2 hold, then*

$$L_n(\beta, c) = \mu(\beta, c)'(\mathcal{I} \otimes Q_{zz})\Delta_n - \frac{1}{2}\mu(\beta, c)'(\mathcal{I} \otimes Q_{zz})\mu(\beta, c) + o_{p_0}(1)$$

for every (β, c), where

$$\Delta_n = \left(\mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell\left(y_{1i} - \gamma_1' x_i, y_{2i} - \gamma_2' x_i\right) \otimes z_i \to_{d_0} \mathcal{N}\left(0, \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right).$$

Theorem 1 is a special case of a local asymptotic normality (LAN) result for the model (3). The general LAN result is given in Theorem A.1 in the Appendix.

As in van der Vaart (1998, Section 9.3), Theorem 1 and Le Cam's third lemma can be used to show that if Assumption 1(a), 2, and 4W hold, then the asymptotically sufficient statistic Δ_n satisfies $\Delta_n \to_d \mathcal{N}\left[\mu(\beta, c), \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right]$, implying in particular that the limiting experiment is that of a single observation from the $\mathcal{N}\left[\mu(\beta, c), \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right]$ distribution whether or not the errors are Gaussian.

Under the same assumptions, the quasi-sufficient (i.e., sufficient when the errors are Gaussian) statistic

$$\bar{\Delta}_n = \left(\Omega \otimes Q_{zz,n}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\ell}\left(y_{1i} - \gamma_1' x_i, y_{2i} - \gamma_2' x_i\right) \otimes z_i,$$

$$\bar{\ell}(v) = \Omega^{-1} v,$$

obtained from the Gaussian quasi-likelihood satisfies $\bar{\Delta}_n \to_d \mathcal{N}\left[\mu(\beta, c), \Omega \otimes Q_{zz}^{-1}\right]$. The Cauchy–Schwarz inequality can be used to show that $\mathcal{I}^{-1} \leq \Omega$, with equality if and only if $\ell(v)$ is linear in v on the support of f. By implication, procedures based on the Gaussian quasi-likelihood are asymptotically inefficient in general. More specifically, any test based on a "smooth" (e.g., almost everywhere continuous) function of $\bar{\Delta}_n$, such as those proposed by Anderson and Rubin (1949), Kleibergen (2002), and Moreira (2003), will be dominated by a test which is efficient (or nearly so) under the assumptions of Theorem 1.[6]

Nevertheless, the results obtained under the assumption of Gaussian errors are of considerable relevance also in models with non-Gaussian errors. This is so because the limiting experiments, indexed by $\mathcal{I}^{-1} \otimes Q_{zz}^{-1}$, in the general case are isomorphic to the limiting experiments, indexed by $\Omega \otimes Q_{zz}^{-1}$, associated with Gaussian errors, a very convenient result because it implies that the insights concerning the relative merits of various testing procedures obtained under the assumption of normality are directly applicable in the general case.

To be specific, let $S_n, T_n \in \mathbb{R}^q$ be given by

$$\left(S_n', T_n'\right)' = \left[\mathcal{I}^{1/2\prime} \otimes Q_{zz}^{1/2\prime}\right]\Delta_n,$$

where $M^{1/2}$ denotes the upper triangular Cholesky factor of a symmetric, positive semi-definite matrix M; that is, $M = M^{1/2}M^{1/2\prime}$, where $M^{1/2}$ is upper triangular.[7] The pair (S_n, T_n) is a non-Gaussian counterpart of $\left(\bar{S}_n', \bar{T}_n'\right)' = \left[\left(\Omega^{-1}\right)^{1/2\prime} \otimes Q_{zz,n}^{1/2\prime}\right]\bar{\Delta}_n$, which features prominently in the work by Moreira (2003), AMS, and others.

In terms of $\left(\bar{S}_n, \bar{T}_n\right)$, the (known Ω) Anderson–Rubin, Lagrange multiplier, and likelihood ratio test statistics popularized by Anderson and Rubin (1949), Kleibergen (2002), and Moreira (2003), respectively, can be expressed as

$$\overline{AR}_n = AR\left(\bar{S}_n\right) = \bar{S}_n'\bar{S}_n,$$

$$\overline{LM}_n = LM\left(\bar{S}_n, \bar{T}_n\right) = \frac{\left(\bar{S}_n'\bar{T}_n\right)^2}{\bar{T}_n'\bar{T}_n},$$

$$\overline{LR}_n = LR\left(\bar{S}_n, \bar{T}_n\right) = \frac{1}{2}\left(\bar{S}_n'\bar{S}_n - \bar{T}_n'\bar{T}_n + \sqrt{\left(\bar{S}_n'\bar{S}_n - \bar{T}_n'\bar{T}_n\right)^2 + 4\left(\bar{S}_n'\bar{T}_n\right)^2}\right).$$

In perfect analogy with the Gaussian case, let $AR_n = AR(S_n)$, $LM_n = LM(S_n, T_n)$, and $LR_n = LR(S_n, T_n)$. The tests which reject H_0 when $AR_n > \chi_\alpha^2(q)$, $LM_n > \chi_\alpha^2(1)$, and $LR_n > \kappa_\alpha(T_n)$ have asymptotic size α, where $\chi_\alpha^2(d)$ is the $1 - \alpha$ quantile of the χ^2 distribution with d degrees of freedom and $\kappa_\alpha(t)$ is the $1 - \alpha$ quantile of the distribution of $LR(\mathcal{Z}, t)$, where $\mathcal{Z} \sim \mathcal{N}\left(0, I_q\right)$.[8] Because of the isomorphism between the Gaussian case and the general case, the relative merits of these testing procedures are well understood from the numerical work of AMS. In particular, it follows from AMS that the test which rejects when $LR_n > \kappa_\alpha(T_n)$ is nearly efficient in the sense that its power function is "close" to the two-sided power envelope for invariant similar tests.

4. Feasible inference procedures

The results of the previous section were obtained under the tacit assumption that γ and f are known. In addition, it was assumed to be known that identification is weak. This section relaxes these assumptions.

4.1. Inference without knowledge of γ and f

First, consider the problem of conducting inference under weak identification without knowledge of the nuisance parameters γ and f. Doing so is easy provided we can find a pair $\left(\hat{\Delta}_n, \hat{\mathcal{I}}_n\right)$ which is asymptotically equivalent to $(\Delta_n, \mathcal{I})$ under weak identification and can be computed without knowledge of (γ, f). To that end, let

$$\hat{\Delta}_n = \left(\hat{\mathcal{I}}_n^{-1} \otimes Q_{zz,n}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\ell}_{i,n} \otimes z_i,$$

$$\hat{\mathcal{I}}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_{i,n}\hat{\ell}_{i,n}',$$

where $\hat{\ell}_{i,n}$ is an estimator of $\ell\left(y_{1i} - \gamma_1' x_i, y_{2i} - \gamma_2' x_i\right)$. In the spirit of Schick (1987), we assume that $\hat{\ell}_{i,n} = \hat{\ell}_n\left(\hat{v}_i\right)$, where $\hat{v}_i = \left(y_{1i} - \hat{\gamma}_{1n}' x_i, y_{2i} - \hat{\gamma}_{2n}' x_i\right)'$ for some estimator $\hat{\gamma}_n = \left(\hat{\gamma}_{1n}', \hat{\gamma}_{2n}'\right)'$ of γ and

$$\hat{\ell}_n(v) = -\frac{\partial \hat{f}_n(v)/\partial v}{\hat{f}_n(v) + a_n}, \qquad \hat{f}_n(v) = \frac{1}{nh_n^2}\sum_{i=1}^{n}K\left(\frac{v - \hat{v}_i}{h_n}\right),$$

[6] Section 4 will exhibit tests which are nearly efficient under the assumptions of Theorem 1.

[7] In particular, letting $\mathcal{I}_{ij}$ denote element (i, j) of $\mathcal{I}$, we have:

$$\mathcal{I}^{1/2} = \begin{pmatrix} \sqrt{\mathcal{I}_{11.2}} & \mathcal{I}_{12}/\sqrt{\mathcal{I}_{22}} \\ 0 & \sqrt{\mathcal{I}_{22}} \end{pmatrix}, \quad \mathcal{I}_{11.2} = \mathcal{I}_{11} - \mathcal{I}_{12}^2/\mathcal{I}_{22}.$$

[8] As shown by Moreira (2003), $\kappa_\alpha(t)$ depends on t only through $\|t\|$, is monotonically decreasing in $\|t\|$, and satisfies $\lim_{\|t\|\to\infty}\kappa_\alpha(t) = \chi_\alpha^2(1)$. The latter result will be utilized when studying the behavior of the test based on LR_n under strong identification.

where K is a kernel and a_n and h_n are positive sequences. Theorem 2 shows that this construction, which does not involve sample splitting, works when the following assumptions hold.

Assumption 5. (a) $K(s_1, s_2) = k(s_1)\,k(s_2)$, where k is a bounded, symmetric, continuously differentiable density with $\int_{\mathbb{R}} r^2 k(r)\,dr + \sup_{r\in\mathbb{R}} |k'(r)|/k(r) < \infty$.

(b) $a_n \to 0$, $h_n \to 0$, and $na_n^2 h_n^4 \to \infty$.

Assumption 6. $\hat{\gamma}_n$ is discrete and $\sqrt{n}\left(\hat{\gamma}_n - \gamma\right) = O_p(1)$.

Remarks. (i) The nonparametric estimation method used here involves two smoothing parameters, h_n and a_n, of which the former is a bandwidth sequence whereas the latter enables us to avoid trimming when handling the density estimator $\hat{f}_n$ appearing in the denominator of $\hat{\ell}_n$.

(ii) If the variances of v_1 and v_2 are suspected to be of different magnitude it may be desirable to let K be a product kernel of the form $K(s_1, s_2) = \sigma_1^{-1}\sigma_2^{-1} k(s_1/\sigma_1)\,k(s_2/\sigma_2)$, where σ_1 and σ_2 are positive constants and k is as in Assumption 5(a). All results, and their proofs, remain valid if Assumption 5(a) is modified in this way.

(iii) Assumption 5(a) holds if k is the logistic density, but not if k is the standard normal density, the reason being that $|k'(r)|/k(r)$ is unbounded when k is the standard normal density. As explained in remark (ii) following the proof of Theorem A.2 in the Appendix, it is possible to accommodate the normal kernel provided the error density f is such that $\dot{f}$ is bounded.

(iv) In Assumption 6, the statement "$\hat{\gamma}_n$ is discrete" is shorthand for the assumption that $\hat{\gamma}_n$ takes only values in the grid $\left\{\varkappa Z/\sqrt{n} \colon Z \in \mathbb{Z}^{2p}\right\}$, where $\varkappa$ is some constant $2p \times 2p$ matrix. Assuming discreteness on the part of an initial estimator is technically convenient and it seems plausible that this assumption can be dropped if additional smoothness is assumed on the part of f. If $\tilde{\gamma}_n$ is a $\sqrt{n}$-consistent estimator of γ, then Assumption 6 will be satisfied by $\hat{\gamma}_n = \left\lfloor \sqrt{n}\tilde{\gamma}_n \right\rfloor / \sqrt{n}$, where $\lfloor \cdot \rfloor$ denotes the integer part of the argument (defined element-by-element). A similar remark applies to Assumptions 7 and 8.

(v) Assumption 6 is satisfied by a discretized version of $\hat{\gamma}_n^{OLS} = \left(\hat{\gamma}_{1,n}^{OLS\prime}, \hat{\gamma}_{2,n}^{OLS\prime}\right)'$, where $\hat{\gamma}_{j,n}^{OLS} = \left(\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\sum_{i=1}^n x_i y_{ji}\right)$ is the OLS estimator of γ_j $(j = 1, 2)$.

Theorem 2. *If Assumptions 1–3, 4W, 5 and 6 hold, then*

$$\left(\hat{\Delta}_n, \hat{\mathcal{I}}_n\right) = (\Delta_n, \mathcal{I}) + o_p(1).$$

In the model (3), the statistic $\hat{\Delta}_n/\sqrt{n}$ can be interpreted as a one-step estimator of δ which uses the zero vector as an initial estimator. As a consequence, Theorem 2 can and will be derived as a special case of a general adaptation result, Theorem A.2 in the Appendix, for one-step estimators of δ in the model (3). Theorem A.2 assumes existence of a discrete $\sqrt{n}$-consistent initial estimator of δ. This requirement is easily met, especially so under weak identification because the zero vector can serve as a $\sqrt{n}$-consistent estimator of δ in that case.[9] Somewhat surprisingly, perhaps, some aspects of conducting inference are therefore simplified by the assumption of weak identification.

Theorem 2 and the continuous mapping theorem can be used to show that if identification is weak, then the local asymptotic power properties of the tests based on AR_n, LM_n, and LR_n are matched by those of the tests based on $\widehat{AR}_n = AR\left(\hat{S}_n\right)$, $\widehat{LM}_n = LM\left(\hat{S}_n, \hat{T}_n\right)$, and $\widehat{LR}_n = LR\left(\hat{S}_n, \hat{T}_n\right)$, respectively, where $\left(\hat{S}_n', \hat{T}_n'\right)' = \left[\hat{\mathcal{I}}_n^{1/2\prime} \otimes Q_{zz,n}^{1/2\prime}\right]\hat{\Delta}_n$. More specifically, we have the following corollary, which implies in particular that the test which rejects when $\widehat{LR}_n > \kappa_\alpha\left(\hat{T}_n\right)$ is nearly efficient when identification is weak.

Corollary 3. *If Assumptions 1–3, 4W, 5 and 6 hold, then*

$$\left[\widehat{AR}_n, \widehat{LM}_n, \widehat{LR}_n, \kappa_\alpha\left(\hat{T}_n\right)\right] = \left[AR_n, LM_n, LR_n, \kappa_\alpha(T_n)\right] + o_p(1).$$

4.2. Inference when identification may be strong

Next, consider the consequences of relaxing the assumption that identification is known to be weak. We are interested in finding a pair of statistics, computable without knowledge of (γ, f), which is asymptotically equivalent to $(\Delta_n, \mathcal{I})$ under weak identification and is "well behaved" also when identification is strong.

When Assumptions 1–3 and 4SC hold, the quasi-sufficient statistic $\bar{\Delta}_n$ obtained from the Gaussian quasi-likelihood satisfies

$$\bar{\Delta}_n - \sqrt{n}\mu(0, \pi) \to_d \mathcal{N}\left[\left(b\pi', 0'\right)', \Omega \otimes Q_{zz}^{-1}\right].$$

It follows immediately from this result that if Assumptions 1–3 and 4SC holds, then

$$\overline{AR}_n \to_d \chi^2\left(q; b^2\pi' Q_{zz}\pi/\omega_{11}\right)$$

and

$$\overline{LM}_n = \overline{LR}_n + o_p(1) = \left(\bar{S}_n' Q_{zz}^{1/2\prime}\pi\right)^2/\pi' Q_{zz}\pi + o_p(1) \to_d \chi^2\left(1; b^2\pi' Q_{zz}\pi/\omega_{11}\right),$$

where ω_{11} is element $(1,1)$ of Ω and $\chi^2(d;\lambda)$ denotes the noncentral χ^2 distribution with d degrees of freedom and noncentrality parameter λ.[10] The convergence result for $\bar{\Delta}_n$ derives in part from the linearity of $\bar{\ell}$ and an analogous result will typically fail to hold for Δ_n and/or $\hat{\Delta}_n$. Indeed, at the present level of generality very little can be said about the asymptotic null properties of statistics such as $\widehat{LR}_n$ under strong identification. This observation motivates the search for a statistic which is asymptotically equivalent to Δ_n under weak identification and exhibits behavior qualitatively similar to that of $\bar{\Delta}_n$ under Assumption 4SC.

Theorem 4 gives conditions under which this property is enjoyed by

$$\hat{\Delta}_n^* = \begin{pmatrix} 0 \\ \sqrt{n}\hat{\pi}_n \end{pmatrix} + \left(\hat{\mathcal{I}}_n^{*-1} \otimes Q_{zz,n}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\ell}_{i,n}^* \otimes z_i,$$

$$\hat{\mathcal{I}}_n^* = \frac{1}{n}\sum_{i=1}^n \hat{\ell}_{i,n}^* \hat{\ell}_{i,n}^{*\prime},$$

with $\hat{\ell}_{i,n}^* = \hat{\ell}_n^*\left(\hat{v}_i^*\right)$, where $\hat{v}_i^* = \left(y_{1i} - \hat{\gamma}_{1n}' x_i, y_{2i} - \hat{\gamma}_{2n}' x_i - \hat{\pi}_n' z_i\right)'$ for some estimators $\left(\hat{\gamma}_n, \hat{\pi}_n\right)$ of (γ, π), and

$$\hat{\ell}_n^*(v) = -\frac{\partial \hat{f}_n^*(v)/\partial v}{\hat{f}_n^*(v) + a_n}, \qquad \hat{f}_n^*(v) = \frac{1}{nh_n^2}\sum_{i=1}^n K\left(\frac{v - \hat{v}_i^*}{h_n}\right).$$

[9] The full force of Theorem A.2 will be needed when Assumption 4W is replaced by Assumption 4SC or 4SF.

[10] Moreover, the properties of κ_α mentioned in footnote 8 at the end of Section 3 can be used to show that $\kappa_\alpha\left(\bar{T}_n\right) = \chi_\alpha^2(1) + o_p(1)$.

As defined, $\hat{\Delta}_n^*/\sqrt{n}$ is a one-step estimator of δ in (3) which uses $\left(0', \hat{\pi}_n'\right)'$ as an initial estimator of δ. This initial estimator is $\sqrt{n}$-consistent under Assumption 4SC provided $\hat{\pi}_n$ satisfies the following condition, which holds if $\hat{\pi}_n$ is a discretized version of $\hat{\pi}_n^{OLS} = \left(\sum_{i=1}^n z_i z_i'\right)^{-1}\left(\sum_{i=1}^n z_i y_{2i}\right)$.

Assumption 7. $\hat{\pi}_n$ is discrete and $\sqrt{n}\left(\hat{\pi}_n - \pi\right) = O_p(1)$.

Theorem 4. (a) *If Assumptions 1–3, 4W and 5–7 hold, then*

$$\left(\hat{\Delta}_n^*, \hat{\mathcal{I}}_n^*\right) = \left(\Delta_n, \mathcal{I}\right) + o_p(1).$$

(b) *If Assumptions 1–3, 4SC and 5–7 hold, then* $\hat{\mathcal{I}}_n^* = \mathcal{I} + o_p(1)$ *and*

$$\hat{\Delta}_n^* - \sqrt{n}\mu(0, \pi) \to_d \mathcal{N}\left[\left(b\pi', 0'\right)', \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right].$$

As a consequence of Theorem 4, we have the following result concerning the statistics $\widehat{AR}_n^* = AR\left(\hat{S}_n^*\right)$, $\widehat{LM}_n^* = LM\left(\hat{S}_n^*, \hat{T}_n^*\right)$, and $\widehat{LR}_n^* = LR\left(\hat{S}_n^*, \hat{T}_n^*\right)$, where $\left(\hat{S}_n^{*\prime}, \hat{T}_n^{*\prime}\right)' = \left[\hat{\mathcal{I}}_n^{*1/2\prime} \otimes Q_{zz,n}^{1/2\prime}\right]\hat{\Delta}_n^*$.

Corollary 5. (a) *If Assumptions 1–3, 4W and 5–7 hold, then*

$$\left[\widehat{AR}_n^*, \widehat{LM}_n^*, \widehat{LR}_n^*, \kappa_\alpha\left(\hat{T}_n^*\right)\right] = \left[AR_n, LM_n, LR_n, \kappa_\alpha(T_n)\right] + o_p(1).$$

(b) *If Assumptions 1–3, 4SC and 5–7 hold, then* $\kappa_\alpha\left(\hat{T}_n^*\right) = \chi_\alpha^2(1) + o_p(1)$ *and*

$$\widehat{AR}_n^* = AR_n + o_p(1) \to_d \chi^2\left(q; b^2\pi' Q_{zz}\pi / \mathcal{I}_{11.2}^{-1}\right),$$
$$\widehat{LM}_n^* = \widehat{LR}_n^* + o_p(1) = \left(S_n' Q_{zz}^{1/2\prime}\pi\right)^2 / \pi' Q_{zz}\pi + o_p(1) \to_d \chi^2\left(1; b^2\pi' Q_{zz}\pi / \mathcal{I}_{11.2}^{-1}\right).$$

It follows from Corollary 5(a) that the test which rejects when $\widehat{LR}_n^* > \kappa_\alpha\left(\hat{T}_n^*\right)$ is nearly efficient when identification is weak. Moreover, Theorem A.1 in the Appendix and Choi et al. (1996, Theorem 2) can be used to show that the test which rejects for large values of $\left(S_n' Q_{zz}^{1/2\prime}\pi\right)^2 / \left(\pi' Q_{zz}\pi\right)$ is asymptotically uniformly most powerful unbiased (in the terminology of Choi et al. (1996, Section 4)) under the assumptions of Corollary 5(b). As a consequence, Corollary 5 (b) implies that the test which rejects when $\widehat{LR}_n^* > \kappa_\alpha\left(\hat{T}_n^*\right)$ enjoys demonstrable optimality properties under strong identification, as does the test which rejects when $\widehat{LM}_n^* > \chi_\alpha^2(1)$. In particular, under strong identification these (asymptotically equivalent) tests are superior to the tests based on the statistics $\overline{AR}_n, \overline{LM}_n, \overline{LR}_n$ and Andrews and Soares (2007) rank-based analogues thereof.

4.3. Consistency

Finally, we address the issue of test consistency under strong identification. The tests based on $\overline{AR}_n$, $\overline{LM}_n$, and $\overline{LR}_n$ are all consistent because $\kappa_\alpha(\cdot)$ is bounded and because

$$n^{-1}\overline{AR}_n = n^{-1}\overline{LM}_n + o_p(1) = n^{-1}\overline{LR}_n + o_p(1)$$
$$= \beta^2\pi' Q_{zz}\pi / \omega_{11} + o_p(1)$$

under Assumptions 1–3 and 4SF, the displayed results following almost immediately from the fact that if Assumptions 1–3 and 4SF hold, then

$$\bar{\Delta}_n - \sqrt{n}\mu(\beta, \pi) \to_d \mathcal{N}\left(0, \Omega \otimes Q_{zz}^{-1}\right).$$

Once again, this convergence result for $\bar{\Delta}_n$ derives in part from the linearity of $\bar{\ell}$ and an analogous result will typically fail to hold for Δ_n, $\hat{\Delta}_n$ and/or $\hat{\Delta}_n^*$. In fact, at the present level of generality there is no guarantee that the tests based on $\widehat{AR}_n^*$, $\widehat{LM}_n^*$, and $\widehat{LR}_n^*$ are consistent under strong identification.

Fortunately this potential problem is easily avoided. Indeed, let

$$\hat{\Delta}_n^{**} = \begin{pmatrix}\sqrt{n}\hat{\Pi}_n \\ \sqrt{n}\hat{\pi}_n\end{pmatrix} + \left(\hat{\mathcal{I}}_n^{**-1} \otimes Q_{zz,n}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\ell}_{i,n}^{**} \otimes z_i,$$
$$\hat{\mathcal{I}}_n^{**} = \frac{1}{n}\sum_{i=1}^n \hat{\ell}_{i,n}^{**}\hat{\ell}_{i,n}^{**\prime},$$

with $\hat{\ell}_{i,n}^{**} = \hat{\ell}_n^{**}(\hat{v}_i^{**})$, where $\hat{v}_i^{**} = (y_{1i} - \hat{\gamma}_{1n}' x_i - \hat{\Pi}_n' z_i, y_{2i} - \hat{\gamma}_{2n}' x_i - \hat{\pi}_n' z_i)'$ for some estimators $(\hat{\gamma}_n, \hat{\pi}_n, \hat{\Pi}_n)$ of $(\gamma, \pi, \beta\pi)$,

$$\hat{\ell}_n^{**}(v) = -\frac{\partial \hat{f}_n^{**}(v)/\partial v}{\hat{f}_n^{**}(v) + a_n}, \qquad \hat{f}_n^{**}(v) = \frac{1}{nh_n^2}\sum_{i=1}^n K\left(\frac{v - \hat{v}_i^{**}}{h_n}\right),$$

and $\hat{\Pi}_n$ is assumed to satisfy the following condition, which holds if $\hat{\Pi}_n$ is a discretized version of $\hat{\Pi}_n^{OLS} = \left(\sum_{i=1}^n z_i z_i'\right)^{-1}\left(\sum_{i=1}^n z_i y_{1i}\right)$.

Assumption 8. $\hat{\Pi}_n$ is discrete and $\sqrt{n}\left(\hat{\Pi}_n - \beta\pi\right) = O_p(1)$.

Once again, $\hat{\Delta}_n^{**}/\sqrt{n}$ can be interpreted as a one-step estimator of δ in (3). Unlike $\hat{\Delta}_n/\sqrt{n}$ and $\hat{\Delta}_n^*/\sqrt{n}$, $\hat{\Delta}_n^{**}/\sqrt{n}$ employs an initial estimator of δ with global $\sqrt{n}$-consistency properties. This feature is utilized in the proof of part (c) of the following result, which in turn can be used to establish consistency of tests based on $\hat{\Delta}_n^{**}$.

Theorem 6. (a) *If Assumptions 1–3, 4W and 5–8 hold, then*

$$\left(\hat{\Delta}_n^{**}, \hat{\mathcal{I}}_n^{**}\right) = \left(\Delta_n, \mathcal{I}\right) + o_p(1).$$

(b) *If Assumptions 1–3, 4SC and 5–8 hold, then* $\hat{\mathcal{I}}_n^{**} = \mathcal{I} + o_p(1)$ *and*

$$\hat{\Delta}_n^{**} - \sqrt{n}\mu(0, \pi) \to_d \mathcal{N}\left[\left(b\pi', 0'\right)', \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right].$$

(c) *If Assumptions 1–3, 4SF and 5–8 hold, then* $\hat{\mathcal{I}}_n^{**} = \mathcal{I} + o_p(1)$ *and*

$$\hat{\Delta}_n^{**} - \sqrt{n}\mu(\beta, \pi) \to_d \mathcal{N}\left(0, \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right).$$

Let $\left(\hat{S}_n^{**\prime}, \hat{T}_n^{**\prime}\right)' = \left[\hat{\mathcal{I}}_n^{**1/2\prime} \otimes Q_{zz,n}^{1/2\prime}\right]\hat{\Delta}_n^{**}$ and define $\widehat{AR}_n^{**} = AR\left(\hat{S}_n^{**}\right)$, $\widehat{LM}_n^{**} = LM\left(\hat{S}_n^{**}, \hat{T}_n^{**}\right)$, and $\widehat{LR}_n^{**} = LR\left(\hat{S}_n^{**}, \hat{T}_n^{**}\right)$. The salient properties of these statistics are characterized in the following corollary to Theorem 6.

Corollary 7. (a) *If Assumptions 1–3, 4W and 5–8 hold, then*

$$\left[\widehat{AR}_n^{**}, \widehat{LM}_n^{**}, \widehat{LR}_n^{**}, \kappa_\alpha\left(\hat{T}_n^{**}\right)\right]$$
$$= \left[AR_n, LM_n, LR_n, \kappa_\alpha(T_n)\right] + o_p(1).$$

(b) *If Assumptions 1–3, 4SC and 5–8 hold, then* $\kappa_\alpha\left(\hat{T}_n^{**}\right) = \chi_\alpha^2(1) + o_p(1)$ *and*

$$\widehat{AR}_n^{**} = AR_n + o_p(1) \to_d \chi^2\left(q; b^2\pi' Q_{zz}\pi / \mathcal{I}_{11.2}^{-1}\right),$$
$$\widehat{LM}_n^{**} = \widehat{LR}_n^{**} + o_p(1) = \left(S_n' Q_{zz}^{1/2\prime}\pi\right)^2 / \pi' Q_{zz}\pi + o_p(1) \to_d \chi^2\left(1; b^2\pi' Q_{zz}\pi / \mathcal{I}_{11.2}^{-1}\right).$$

(c) *If Assumptions 1–3, 4SF and 5–8 hold, then*

$$\begin{aligned} n^{-1}\widehat{AR}_n^{**} &= n^{-1}\widehat{LM}_n^{**} + o_p(1) = n^{-1}\widehat{LR}_n^{**} + o_p(1) \\ &= \beta^2\pi' Q_{zz}\pi / \mathcal{I}_{11.2}^{-1} + o_p(1). \end{aligned}$$

In perfect analogy with Corollary 5, parts (a) and (b) of Corollary 7 imply that the test which rejects when $\widehat{LR}_n^{**} > \kappa_\alpha\left(\hat{T}_n^{**}\right)$ is nearly optimal when identification is weak and demonstrably optimal when identification is strong. Relative to Corollary 5, which establishes analogous results for the test which rejects when $\widehat{LR}_n^{*} > \kappa_\alpha\left(\hat{T}_n^{*}\right)$, the additional property that can be claimed on the part of the test based on $\widehat{LR}_n^{**}$ is that of consistency under strong identification. This, and the analogous consistency results about the tests based on $\widehat{AR}_n^{**}$ and $\widehat{LM}_n^{**}$, is the content of Corollary 7(c).

4.4. Inference when identification is strong

If identification is strong, then the usual duality between estimation and testing holds, implying in particular that the asymptotic optimality properties of the tests based on $\widehat{LR}_n^{**}$ and $\widehat{LM}_n^{**}$ are shared by a Wald test based on an asymptotically efficient estimator of β.

Let

$$\hat{\beta}_n^{**} = \hat{\Delta}_{1,n}^{**\prime} Q_{zz,n}\hat{\Delta}_{2,n}^{**} / \hat{\Delta}_{2,n}^{**\prime} Q_{zz,n}\hat{\Delta}_{2,n}^{**},$$

where $\hat{\Delta}_n^{**} = \left(\hat{\Delta}_{1,n}^{**\prime}, \hat{\Delta}_{2,n}^{**\prime}\right)'$ and partitioning is after the qth row. The estimator $\hat{\beta}_n^{**}$ can be interpreted as a non-Gaussian counterpart of the 2SLS estimator of β, the latter being given by $\bar{\beta}_n = \bar{\Delta}_{1,n}' Q_{zz,n}\bar{\Delta}_{2,n} / \bar{\Delta}_{2,n}' Q_{zz,n}\bar{\Delta}_{2,n}$, where $\bar{\Delta}_n = \left(\bar{\Delta}_{1,n}', \bar{\Delta}_{2,n}'\right)'$. The estimators $\hat{\beta}_n^{**}$ and $\bar{\beta}_n$ are both obtained by means of a generalized least squares (GLS) regression of an estimator of δ_1 onto an estimator of δ_2 in (3). The GLS regressions utilize identical weighting matrices, but differ in terms of the estimators of δ being employed, with $\hat{\beta}_n^{**}$ being based on an asymptotically efficient estimator, $\hat{\Delta}_n^{**}/\sqrt{n}$, and $\bar{\beta}_n$ being based on the OLS estimator $\bar{\Delta}_n/\sqrt{n}$.

If Assumptions 1–3 and 4SF hold, then $\sqrt{n}\left(\bar{\beta}_n - \beta\right) \to_d \mathcal{N}\left(0, \bar{\Sigma}_\beta\right)$, where $\bar{\Sigma}_\beta = \left[(1, -\beta)\,\Omega\,(1, -\beta)'\right] / \pi' Q_{zz}\pi$. The next result, which follows from Theorem 6(c) and the delta method, gives the corresponding result for $\hat{\beta}_n^{**}$.

Corollary 8. *If Assumptions 1–3, 4SF and 5–8 hold, then*

$$\begin{aligned} &\sqrt{n}\left(\hat{\beta}_n^{**} - \beta\right) \to_d \mathcal{N}\left(0, \Sigma_\beta\right), \\ &\quad \Sigma_\beta = \left[(1, -\beta)\,\mathcal{I}^{-1}\,(1, -\beta)'\right] / \pi' Q_{zz}\pi. \end{aligned}$$

Under normality the convergence result in Corollary 8 agrees with that for the 2SLS estimator of β and its asymptotic equivalents, such as the limited information maximum likelihood (LIML) estimator and Fuller's (1977) modification thereof. With non-Gaussian errors, on the other hand, the estimator $\hat{\beta}_n^{**}$ compares favorably with $\bar{\beta}_n$ whenever the inequality $\mathcal{I}^{-1} \le \Omega$ is strict.

The existence of estimators which outperform 2SLS for certain non-Gaussian error distributions has been known at least since Amemiya (1982) and Powell (1983). For the purposes of relating $\hat{\beta}_n^{**}$ to the two-stage least absolute deviations (2SLAD) and double 2SLAD (D2SLAD) estimators studied in those papers, define

$$\begin{aligned} \tilde{\beta}_n(\lambda_1, \lambda_2) = {} & \hat{\Pi}_n(\lambda_1)' Q_{zz,n}\hat{\pi}_n(\lambda_2) / \hat{\pi}_n(\lambda_2)' \\ & \times Q_{zz,n}\hat{\pi}_n(\lambda_2), \quad (\lambda_1, \lambda_2)' \in \mathbb{R}^2, \end{aligned}$$

where $\hat{\Pi}_n(\lambda_1) = \lambda_1\hat{\Pi}_n^{LAD} + (1-\lambda_1)\,\hat{\Pi}_n^{OLS}$, $\hat{\pi}_n(\lambda_2) = \lambda_2\hat{\pi}_n^{LAD} + (1-\lambda_2)\,\hat{\pi}_n^{OLS}$, and

$$\begin{aligned} \left(\hat{\Pi}_n^{LAD}, \hat{\pi}_n^{LAD}\right) = \arg\min_{(\Pi,\pi)}\min_{(\gamma_1,\gamma_2)} \sum_{i=1}^{n} & \left|y_{1i} - \gamma_1' x_i - \Pi' z_i\right| \\ & + \left|y_{2i} - \gamma_2' x_i - \pi' z_i\right|. \end{aligned}$$

In this notation $\tilde{\beta}_n(0, 0)$ is the 2SLS estimator, while nonzero pairs (λ_1, λ_2) give rise to estimators that are asymptotically distinct from the 2SLS estimator. The Bahadur representation of any $\tilde{\beta}_n(\lambda_1, \lambda_2)$ is readily obtained from the Bahadur representations of $\hat{\Pi}_n^{LAD}$, $\hat{\Pi}_n^{OLS}$, $\hat{\pi}_n^{LAD}$, and $\hat{\pi}_n^{OLS}$. Utilizing these Bahadur representations it can be shown that $\tilde{\beta}_n(\lambda_1, 0)$ is asymptotically equivalent to the 2SLAD(λ_1) estimator and that $\tilde{\beta}_n(1, 1)$ is asymptotically equivalent to the D2SLAD estimator(s).

Because $\left(\hat{\Delta}_{1,n}^{**}, \hat{\Delta}_{2,n}^{**}\right)/\sqrt{n}$ is an asymptotically efficient estimator of (δ_1, δ_2) in (3), it compares favorably with $\left(\hat{\Pi}_n(\lambda_1), \hat{\pi}_n(\lambda_2)\right)$ for any value of (λ_1, λ_2). This superiority is inherited by $\hat{\beta}_n^{**}$, which compares favorably with all estimators of the form $\tilde{\beta}_n(\lambda_1, \lambda_2)$ (and their asymptotic equivalents, such as the 2SLAD and D2SLAD estimators). In fact, Theorems A.1 and A.2 can be used to show that $\hat{\beta}_n^{**}$ is an asymptotically efficient (i.e., best regular) estimator of β under strong identification.

As a consequence, one would expect the strong identification local asymptotic power properties of the tests based on $\widehat{LR}_n^{**}$ and $\widehat{LM}_n^{**}$ to be matched by those of the test which rejects when $\widehat{W}_n^{**} > \chi_\alpha^2(1)$, where

$$\begin{aligned} \widehat{W}_n^{**} &= n\left(\hat{\beta}_n^{**}\right)^2 / \hat{\Sigma}_\beta^{**}, \\ \hat{\Sigma}_\beta^{**} &= \left[\left(1, -\hat{\beta}_n^{**}\right)\hat{\mathcal{I}}_n^{**-1}\left(1, -\hat{\beta}_n^{**}\right)'\right] / \hat{\pi}_n' Q_{zz,n}\hat{\pi}_n. \end{aligned}$$

The next result, which follows from Theorem 6 (b) and the delta method, verifies that conjecture.

Corollary 9. *If Assumptions 1–3, 4SC and 5–8 hold, then*

$$\begin{aligned} \widehat{W}_n^{**} &= \left(S_n' Q_{zz}^{1/2\prime}\pi\right)^2 / \pi' Q_{zz}\pi + o_p(1) \\ &\to_d \chi^2\left(1; b^2\pi' Q_{zz}\pi / \mathcal{I}_{11.2}^{-1}\right). \end{aligned}$$

An attractive feature of $\widehat{W}_n^{**}$ is that its ingredients, $\hat{\beta}_n^{**}$ and $\hat{\Sigma}_\beta^{**}$, can be combined in the usual way to form a Wald test of any null hypothesis regarding β, not just the null hypothesis that $\beta = 0$. This feature is particularly convenient when hypothesis tests are used to construct confidence intervals by inversion, as it implies that valid (indeed, optimal) confidence intervals are trivial to construct. Indeed, a confidence interval with asymptotic coverage probability $1-\alpha$ is given by

$$\left(\hat{\beta}_n^{**} - \sqrt{\chi_\alpha^2(1)\,\hat{\Sigma}_\beta^{**}/n},\ \hat{\beta}_n^{**} + \sqrt{\chi_\alpha^2(1)\,\hat{\Sigma}_\beta^{**}/n}\right).$$

It should be emphasized, however, that the displayed confidence interval does not have asymptotic coverage probability $1-\alpha$ under weak identification. As a consequence, while the computational simplicity of $\widehat{W}_n^{**}$ makes it an attractive competitor to $\widehat{LM}_n^{**}$ and $\widehat{LR}_n^{**}$ under strong identification, the Wald statistic does not enjoy the robustness (and, in the case of $\widehat{LR}_n^{**}$, near optimality) properties under weak identification that Corollary 7 (a) establishes on the part of $\widehat{LM}_n^{**}$ and $\widehat{LR}_n^{**}$.

Remark. The LIMLK (i.e., LIML with known Ω) estimator of β is given by

$$\arg\min_{\beta} \frac{(1,-\beta)\left(\bar{\Delta}_{1,n},\bar{\Delta}_{2,n}\right)' Q_{zz,n}\left(\bar{\Delta}_{1,n},\bar{\Delta}_{2,n}\right)(1,-\beta)'}{(1,-\beta)\,\Omega\,(1,-\beta)'}.$$

This estimator is asymptotically equivalent to the 2SLS estimator $\bar{\beta}_n$ when identification is strong, but enjoys certain advantages over $\bar{\beta}_n$ when identification is weak (e.g., Staiger and Stock (1997)). Analogously, the following non-Gaussian counterpart of the LIMLK estimator of β is asymptotically equivalent (superior) to $\hat{\beta}_n^{**}$ under strong (weak) identification:

$$\arg\min_{\beta} \frac{(1,-\beta)\left(\hat{\Delta}^{**}_{1,n},\hat{\Delta}^{**}_{2,n}\right)' Q_{zz,n}\left(\hat{\Delta}^{**}_{1,n},\hat{\Delta}^{**}_{2,n}\right)(1,-\beta)'}{(1,-\beta)\,\hat{\mathcal{I}}_n^{**-1}\,(1,-\beta)'}.$$

5. Simulations

This section presents the results of a simulation study investigating the finite-sample performance of the procedure considered in this paper. Although we primarily focus on power properties of the tests based on $\widehat{AR}_n^{**}$, $\widehat{LM}_n^{**}$, and $\widehat{LR}_n^{**}$, we also discuss the properties of the point estimator $\hat{\beta}_n^{**}$ under strong identification.

5.1. Model setup

The data are generated by the model (2). Specifically, we set $x_i = 1$ and set q, the dimension of the instrumental variable, equal to 4. The instruments are randomly generated from a standard Gaussian distribution, demeaned, and then kept fixed throughout the experiment. For the errors we consider two different specifications, based on (i) the standard normal distribution and (ii) the $t(3)$ distribution, respectively. (The Fisher information for the location model generated by the $t(3)$ distribution is $2/3$, twice the inverse of the variance of the $t(3)$ distribution.) The probability densities associated with the distributions are depicted in Fig. 1.

We generate $2n$ independent (studentized) errors $\tilde{v}_i = (\tilde{v}_{1i}, \tilde{v}_{2i})'$ from each distribution and define $(v_{1i}, v_{2i})' = (\tilde{v}_{1i}, \sqrt{1-\rho^2}\tilde{v}_{2i} + \rho\tilde{v}_{1i})'$, hereby inducing a correlation of ρ between the errors v_{1i} and v_{2i}. Consistent with the previous discussion, we take $\beta_0 = 0$. The 4×1 vector π is given by $\iota \cdot \sqrt{\zeta q}/\sqrt{\iota' Z' Z\iota}$, where ι is a 4×1 vector of ones, Z is the $n \times 4$ matrix of instruments, and ζ is the concentration parameter $\pi' Z' Z\pi/q$, which determines the "strength" of the instruments. For the simulations, we chose $n = 1000$ as the sample size, $S = 5000$ as the number of simulations, $\rho = 0.5$, and ζ taking on the values 1 and 10. In addition we chose $\alpha = 0.05$ for the size of our tests. (We obtained qualitatively similar results for other choices of n, S, ρ, and ζ, but omit these to conserve space.)

5.2. Implementation

The new procedures are compared to three benchmark procedures. The first of these is the Gaussian procedure constructed using a feasible version of the quasi-sufficient statistics $\left(\bar{S}_n, \bar{T}_n\right)$ employing the OLS estimator $\hat{\Omega}_n^{OLS} = (n-5)^{-1}\sum_{i=1}^n \hat{v}_i^{OLS}\hat{v}_i^{OLS\prime}$ of Ω, where $\hat{v}_i^{OLS}$ are the OLS residuals. We will refer to this technique as "OLS" for simplicity.

As a second benchmark procedure we compute the Normal Scores Rank Tests introduced by Andrews and Soares (2007). We refer to this procedure as "RNK" for brevity. These tests are seen to have superior power properties to those denoted OLS herein and are recommended by the authors based on both asymptotic and finite-sample results. However, based on our asymptotic results, our procedures are expected to have superior power properties over the corresponding RNK tests.

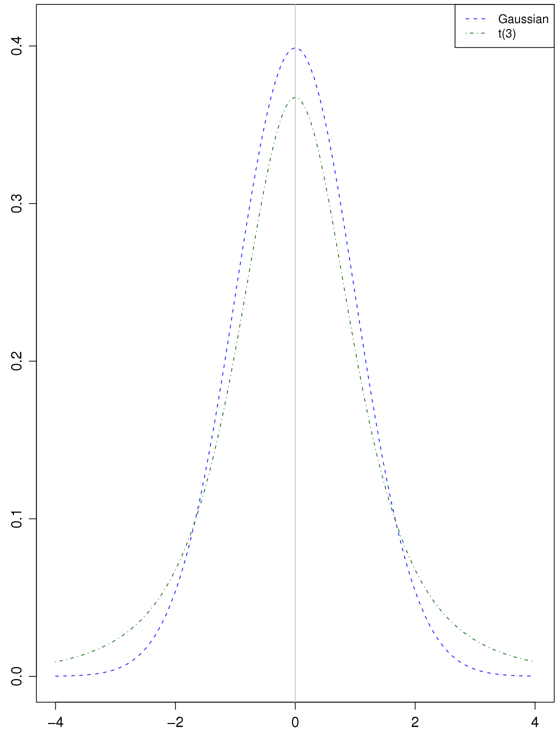

Fig. 1. Probability densities.

Finally, the third benchmark procedure utilizes an "oracle" version of $\hat{\Delta}_n^{**}$. Specifically, using the true ℓ instead of its estimate, we obtain

$$\hat{\Delta}_n^{MLE} = \begin{pmatrix}\sqrt{n}\hat{\Pi}_n^{OLS}\\ \sqrt{n}\hat{\pi}_n^{OLS}\end{pmatrix} + \left(\hat{\mathcal{I}}_\ell^{-1} \otimes Q_{zz,n}^{-1}\right) \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \ell\left(\hat{v}_i^{OLS}\right) \otimes z_i,$$

where $\hat{\mathcal{I}}_\ell = n^{-1}\sum_{i=1}^n \ell\left(\hat{v}_i^{OLS}\right)\ell\left(\hat{v}_i^{OLS}\right)'$. It should be noted that this is not a true "oracle" procedure in the sense that it uses the estimated error terms rather than their true values and also relies on an estimate of the information matrix. We include this additional benchmark in an effort to identify the effects on performance of using nonparametric estimates of the score function. Although a slight abuse of notation, we will refer to this technique as "MLE" for simplicity.

The (feasible) adaptive procedure based on $\hat{\Delta}_n^{**}$ is referred to as "ADP" for notational simplicity. This procedure is fully data-driven, but requires the additional choice of three parameters: the kernel k, the trimming parameter a, and the smoothing parameter h. For specificity we set k equal to a standard Gaussian kernel and set $a = 0$. The choice $a = 0$ violates Assumption 5(b), but was made for simplicity and concreteness because the qualitative results seemed to be more sensitive to the choice of h than to the choice of a. Regarding the choice of h, we experimented with a variety of procedures and specifications. In terms of procedures we considered both first-generation and second-generation bandwidth selection procedures for both univariate density and derivative estimation and bivariate density and derivative estimation (e.g., Ichimura and Todd (2007)). In terms of specifications, we considered a common bandwidth as well as different combinations of alternative bandwidths for densities and partial derivatives. Unfortunately, but unsurprisingly in light of previous Monte Carlo results on adaptive estimation in the

Table 1
Empirical size-$n = 500$.

Bandwidth	$v_i \sim \mathcal{N}(0, 1) - \zeta = 1$			$v_i \sim \mathcal{N}(0, 1) - \zeta = 10$			$v_i \sim t(3) - \zeta = 1$			$v_i \sim t(3) - \zeta = 10$		
Scale (c)	AR	LM	LR	AR	LM	LR	AR	LM	LR	AR	LM	LR
0.45	0.187	0.132	0.173	0.182	0.117	0.133	0.078	0.072	0.080	0.080	0.071	0.077
0.46	0.172	0.126	0.164	0.170	0.111	0.128	0.073	0.068	0.075	0.073	0.067	0.074
0.47	0.159	0.119	0.156	0.161	0.105	0.120	0.067	0.064	0.069	0.068	0.063	0.069
0.48	0.150	0.114	0.147	0.149	0.100	0.115	0.061	0.061	0.066	0.062	0.060	0.066
0.49	0.141	0.109	0.136	0.138	0.095	0.110	0.057	0.059	0.063	0.058	0.057	0.062
0.50	0.132	0.105	0.128	0.131	0.091	0.103	0.053	0.056	0.059	0.055	0.055	0.061
0.51	0.124	0.100	0.120	0.121	0.088	0.099	0.049	0.054	0.056	0.051	0.053	0.058
0.52	0.113	0.094	0.114	0.112	0.084	0.095	0.047	0.051	0.053	0.047	0.050	0.056
0.53	0.106	0.090	0.107	0.107	0.080	0.091	0.044	0.049	0.050	0.045	0.048	0.053
0.54	0.099	0.087	0.099	0.100	0.075	0.088	0.041	0.047	0.048	0.042	0.047	0.051
0.55	0.092	0.083	0.094	0.095	0.072	0.085	0.039	0.044	0.045	0.040	0.045	0.050
0.56	0.084	0.079	0.089	0.086	0.069	0.082	0.038	0.043	0.042	0.038	0.043	0.049
0.57	0.080	0.076	0.084	0.079	0.066	0.078	0.036	0.042	0.040	0.035	0.041	0.046
0.58	0.076	0.071	0.078	0.072	0.064	0.075	0.034	0.040	0.038	0.033	0.039	0.045
0.59	0.071	0.067	0.073	0.069	0.061	0.072	0.032	0.038	0.037	0.031	0.038	0.044
0.60	0.067	0.064	0.069	0.063	0.059	0.068	0.029	0.037	0.034	0.029	0.037	0.042
0.61	0.063	0.062	0.065	0.059	0.056	0.065	0.028	0.036	0.033	0.027	0.036	0.041
0.62	0.059	0.060	0.060	0.055	0.053	0.062	0.027	0.034	0.031	0.025	0.035	0.039
0.63	0.055	0.057	0.057	0.052	0.051	0.059	0.025	0.033	0.029	0.024	0.032	0.038
0.64	0.053	0.056	0.054	0.047	0.049	0.057	0.024	0.032	0.028	0.023	0.030	0.036
0.65	0.049	0.054	0.052	0.045	0.047	0.055	0.021	0.031	0.028	0.021	0.029	0.035
0.66	0.045	0.053	0.050	0.044	0.046	0.053	0.020	0.030	0.027	0.020	0.028	0.033
0.67	0.042	0.051	0.048	0.041	0.045	0.050	0.019	0.028	0.025	0.019	0.028	0.032
0.68	0.039	0.050	0.045	0.039	0.043	0.048	0.018	0.027	0.024	0.018	0.027	0.032
0.69	0.036	0.048	0.043	0.037	0.041	0.046	0.017	0.026	0.024	0.018	0.026	0.030
0.70	0.034	0.045	0.041	0.034	0.039	0.045	0.016	0.026	0.022	0.017	0.025	0.030
0.71	0.033	0.044	0.038	0.030	0.038	0.043	0.016	0.024	0.020	0.017	0.024	0.030
0.72	0.031	0.043	0.037	0.028	0.038	0.043	0.015	0.024	0.020	0.016	0.024	0.029
0.73	0.028	0.041	0.035	0.028	0.037	0.041	0.014	0.024	0.019	0.015	0.023	0.028
0.74	0.027	0.040	0.033	0.025	0.036	0.040	0.014	0.023	0.018	0.014	0.023	0.027
0.75	0.025	0.039	0.031	0.024	0.035	0.039	0.014	0.021	0.017	0.013	0.021	0.026
0.76	0.024	0.037	0.030	0.022	0.034	0.038	0.014	0.021	0.017	0.012	0.020	0.024
0.77	0.023	0.037	0.028	0.021	0.033	0.037	0.013	0.020	0.015	0.012	0.020	0.023
0.78	0.022	0.035	0.026	0.020	0.031	0.035	0.012	0.020	0.015	0.012	0.019	0.022
0.79	0.020	0.034	0.025	0.019	0.030	0.034	0.012	0.019	0.015	0.011	0.019	0.021
0.80	0.019	0.034	0.023	0.018	0.030	0.034	0.011	0.018	0.014	0.011	0.019	0.021

univariate case (e.g., Steigerwald (1992)), our preliminary findings showed that these procedures have disappointing size properties for modest sample sizes. In the end we therefore opted for a simple re-scaling of a rule of thumb choice for bivariate density estimation. Specifically, we set $h_j = c\sqrt{\hat{\omega}_{jj}^{OLS}}n^{-\lambda}$ $(j = 1, 2)$, where h_1 and h_2 are the bandwidth choice for the first and second dimension of the nonparametric score estimator, respectively, and c and λ are constants to be chosen. Regarding λ, the goal was to achieve size distortions that exhibit minimal sensitivity with respect to sample size. The choice $\lambda = 1/9$ was found to be satisfactory and was employed when constructing Tables 1–3, which report size as a function of c for each of the four designs considered and for samples of size $n = 500$ (Table 1), $n = 1000$ (Table 2), and $n = 5000$ (Table 3), respectively. In all cases, size is a decreasing function of c. The size properties are somewhat sensitive to the true distribution of the error terms. Although there does not seem to be any choice of c for which size distortions are close to zero across all designs, the choices $c = 0.65$ and $c = 0.50$ seem to work well for the Gaussian and $t\ (3)$ model, respectively.

5.3. Results

Our explicit goal is to explore the extent to which the asymptotic optimality properties of adaptive procedures are inherited at least partially in finite samples, we report power graphs based on choices of h that deliver tests with actual size close to nominal size in our simulations. (Power curves are easier to interpret and compare when competing tests have common size.) Accordingly, based on the findings reported in Tables 1–3 we set $c = 0.65$ and $c = 0.50$ for the Gaussian and $t\ (3)$ model, respectively. Fig. 2 presents the power graphs for the AR, LM, and CLR tests for the case where the reduced form errors are generated from a Gaussian distribution.

The strength of the instruments is equal to 1 and 10 in the first and second rows of graphs, respectively. In this particular case, the OLS and MLE estimators of the linear coefficients coincide, while the second-moment matrices are equal up to a constant multiple which converges to 1 with the sample size. As a consequence, the power curves of the tests based on these two procedures are virtually equivalent. The RNK tests also appear to reach the power curve generated by OLS and MLE. Because the adaptive procedures employ a nonparametric estimator of ℓ, we would expect them to have reduced finite sample power relative to the "oracle" procedures and this does indeed seem to be the case. Nevertheless, the power loss is encouragingly small and the findings suggest that the ADP procedures can dominate the OLS and RNK procedures when the errors are non-Gaussian.

Fig. 3 presents the results for the case when the errors are generated from a non-Gaussian distribution, a $t(3)$ distribution in this case. Again, the first and second rows differ by the choice of the strength of the instruments. The results in this case are consistent with the theoretical predictions. The tests based on the MLE estimator have superior power relative to the test statistics based on the OLS estimator, while the test statistics based on the RNK procedure and ADP estimator have power curves which reside in between the other two. Moreover, tests based on ADP appear to (non-strictly) dominate the corresponding RNK tests. As expected, the MLE estimator delivers important power gains when compared

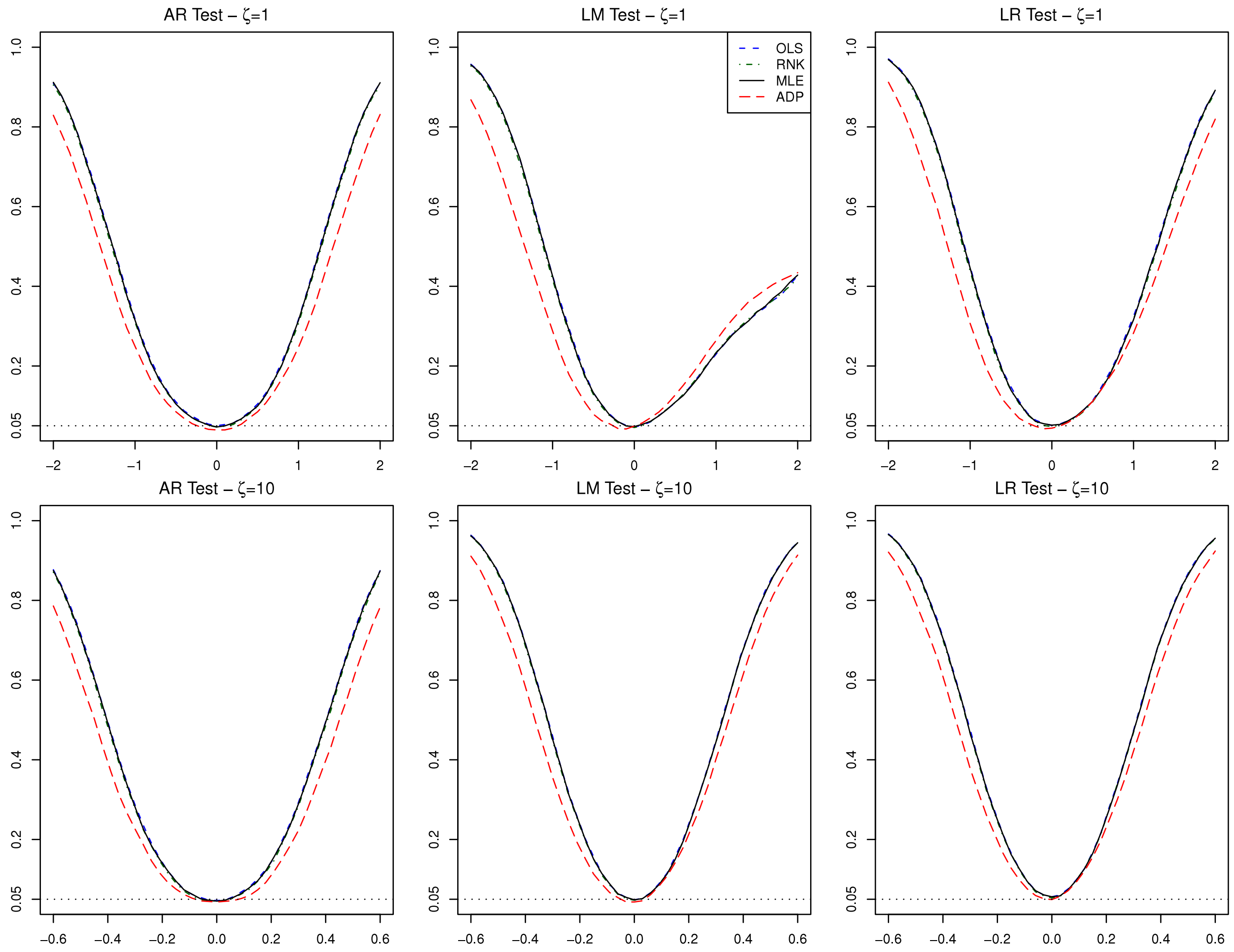

Fig. 2. Power curves, Gaussian errors.

to the RNK procedure. Presumably the difference between the MLE and ADP power curves can be attributed to the fact that ADP employs a nonparametric estimator of the nuisance parameter ℓ. In other words, the asymptotic theory probably overstates the extent to which departures from Gaussianity can be exploited in finite samples. On the other hand, the qualitative predictions of the asymptotic theory are borne out in the simulations insofar as Fig. 3 clearly suggests that even in finite samples the ADP procedures can enjoy power advantages over the OLS procedures when the errors are non-Gaussian.

Finally, in Fig. 4 we present (kernel density estimators of) the sampling distributions of the estimators of β using each procedure when instruments are "strong". The sampling distribution of the ADP estimator $\hat{\beta}_n^{**}$ is more concentrated than that of the "OLS" estimator $\bar{\beta}_n$ and less concentrated than that of the "oracle" estimator. This is also consistent with the theoretical predictions. (Similar results were obtained for the ADP estimator of the reduced-form coefficients. We omit the results to conserve space.)

In our view, the Monte Carlo results provide evidence in favor of the procedure(s) developed in this paper. The key potential drawback of the new procedure(s), which is common to all nonparametric procedures, is the fact that no firm guidance on the choice of the smoothing parameter is available. Although it is beyond the scope of this paper to develop a theory-based bandwidth selection rule with uniformly good size (and power) properties, our results lead us to recommend the use of bandwidths of the form $h_j = 0.65\sqrt{\hat{\omega}_{jj}^{OLS}}n^{-1/9}$, as this choice yields good results when the errors are Gaussian and seems to be conservative otherwise.

Appendix. Proofs

The main results of the paper will follow from two facts, Theorems A.1 and A.2, about the model (3). Neither result is particularly surprising, but we have been unable to find statements of these results in the literature.

Theorem A.1 is an LAN result. To state it, let

$$\begin{aligned}\mathcal{L}_n(d,g) = \sum_{i=1}^{n} \log f \big[& y_{1i} - \gamma_{1n}(g_1)' x_i \\ & - \delta_{1n}(d_1)' z_i, y_{2i} - \gamma_{2n}(g_2)' x_i - \delta_{2n}(d_2)' z_i \big] \\ & - \sum_{i=1}^{n} \log f \left[y_{1i} - \gamma_1' x_i - \delta_1' z_i, y_{2i} - \gamma_2' x_i - \delta_2' z_i \right]\end{aligned}$$

denote the log likelihood ratio function associated with the local reparameterization

$$\gamma = \begin{bmatrix} \gamma_{1n}(g_1) \\ \gamma_{2n}(g_2) \end{bmatrix} = \begin{bmatrix} \gamma_1 + g_1/\sqrt{n} \\ \gamma_2 + g_2/\sqrt{n} \end{bmatrix},$$

$$\delta = \begin{bmatrix} \delta_{1n}(d_1) \\ \delta_{2n}(d_2) \end{bmatrix} = \begin{bmatrix} \delta_1 + d_1/\sqrt{n} \\ \delta_2 + d_2/\sqrt{n} \end{bmatrix},$$

let "$o_{p_{\delta,\gamma}}(1)$" and "$\rightarrow_{d_{\delta,\gamma}}$" be shorthand for "$o_p(1)$ under the distributions associated with $(d,g) = (0,0)$" and "$\rightarrow_d$ under the

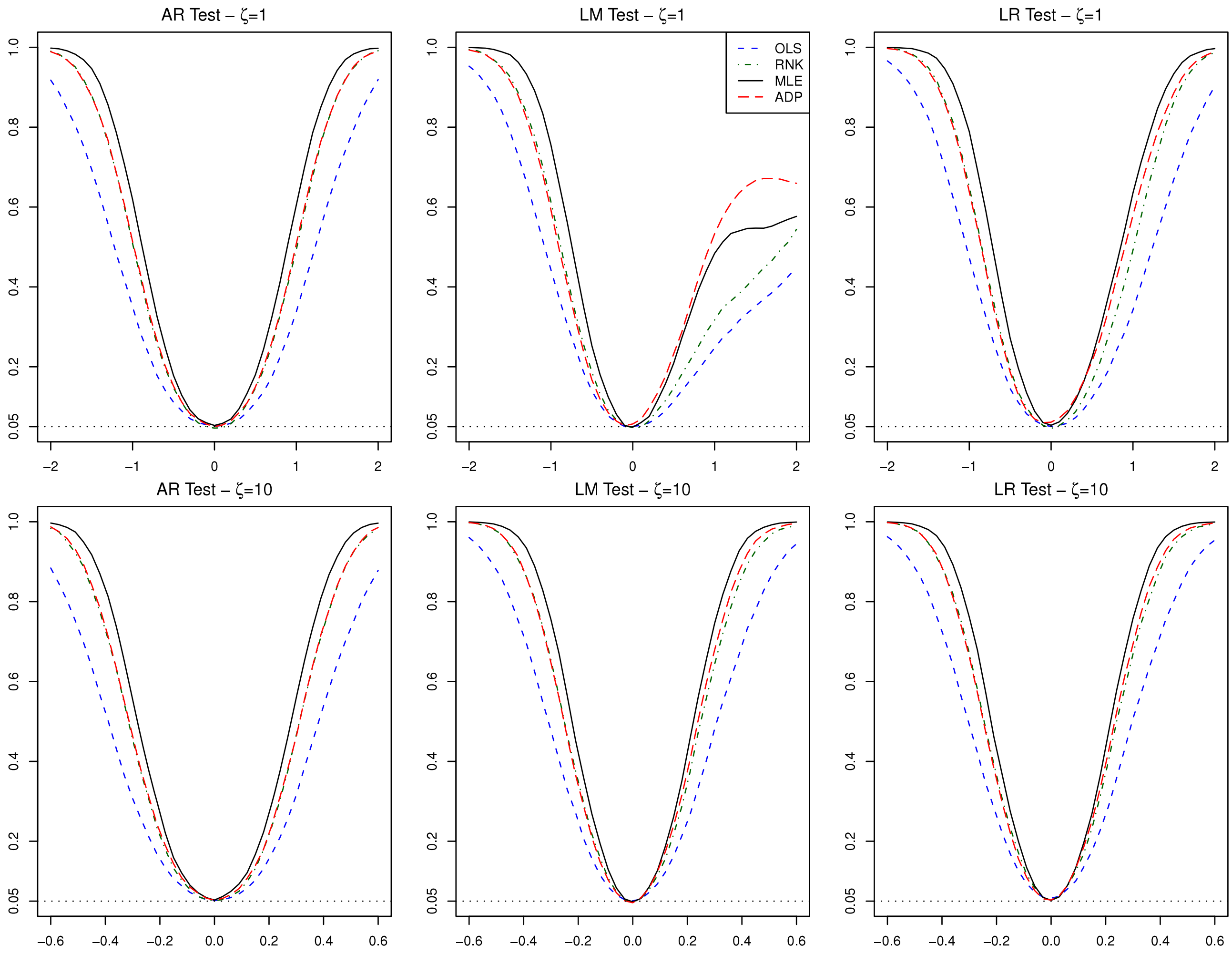

Fig. 3. Power curves, non-Gaussian errors.

distributions associated with $(d, g) = (0, 0)$", respectively, and let $\ell_i = \ell\left(y_{1i} - \gamma_1' x_i - \delta_1' z_i, y_{2i} - \gamma_2' x_i - \delta_2' z_i\right)$.

Theorem A.1. *Suppose (y_{1i}, y_{2i}) is generated by (3).*

(a) *If Assumption 1 (a) and 2 hold and d_n is a bounded sequence, then*

$$\mathcal{L}_n(d_n, 0) = \mathcal{L}_n^{\delta}(d_n) + o_{p_{\delta,\gamma}}(1),$$

where $\mathcal{L}_n^{\delta}(d_n) = d_n'(\mathcal{I} \otimes Q_{zz})\Delta_n^{\delta} - d_n'(\mathcal{I} \otimes Q_{zz})d_n/2$ *and*

$$\Delta_n^{\delta} = \left(\mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_i \otimes z_i \to_{d_{\delta,\gamma}} \mathcal{N}\left(0, \mathcal{I}^{-1} \otimes Q_{zz}^{-1}\right).$$

(b) *If, moreover, Assumption 1 (b) and 3 hold and g_n is a bounded sequence, then*

$$\mathcal{L}_n(d_n, g_n) = \mathcal{L}_n^{\delta}(d_n) + \mathcal{L}_n^{\gamma}(g_n) + o_{p_{\delta,\gamma}}(1),$$

where $\mathcal{L}_n^{\gamma}(g_n) = g_n'(\mathcal{I} \otimes Q_{xx})\Delta_n^{\gamma} - g_n'(\mathcal{I} \otimes Q_{xx})g_n/2$ *and*

$$\begin{pmatrix}\Delta_n^{\delta}\\ \Delta_n^{\gamma}\end{pmatrix} \to_{d_{\delta,\gamma}} \mathcal{N}\left[0, \mathcal{I}^{-1} \otimes \begin{pmatrix}Q_{zz}^{-1} & 0\\ 0 & Q_{xx}^{-1}\end{pmatrix}\right],$$

$$\Delta_n^{\gamma} = \left(\mathcal{I}^{-1} \otimes Q_{xx}^{-1}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_i \otimes x_i.$$

Theorem A.2 is an adaptation result for one-step estimators of δ. Given initial estimators $\hat{\delta}_n = \left(\hat{\delta}_{1n}', \hat{\delta}_{2n}'\right)'$ and $\hat{\gamma}_n = \left(\hat{\gamma}_{1n}', \hat{\gamma}_{2n}'\right)'$ of δ and γ, let

$$\tilde{\delta}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right) = \hat{\delta}_n + \left[\tilde{\mathcal{I}}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right)^{-1} \otimes Q_{zz,n}^{-1}\right]\frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_n\left(\hat{v}_i\right) \otimes z_i,$$

where

$$\tilde{\mathcal{I}}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_n\left(\hat{v}_i\right)\hat{\ell}_n\left(\hat{v}_i\right)',$$

$$\hat{v}_i = \begin{pmatrix}y_{1i} - \hat{\gamma}_{1n}' x_i - \hat{\delta}_{1n}' z_i\\ y_{2i} - \hat{\gamma}_{2n}' x_i - \hat{\delta}_{2n}' z_i\end{pmatrix},$$

$$\hat{\ell}_n(v) = -\frac{\partial \hat{f}_n(v)/\partial v}{\hat{f}_n(v) + a_n}, \quad \hat{f}_n(v) = \frac{1}{nh_n^2}\sum_{i=1}^{n}K\left(\frac{v - \hat{v}_i}{h_n}\right).$$

Theorem A.2. *Suppose (y_{1i}, y_{2i}) is generated by (3). If Assumptions 1–3 and 5 hold, $\left(\hat{\delta}_n, \hat{\gamma}_n\right)$ is discrete, and $\sqrt{n}\left(\hat{\delta}_n - \delta, \hat{\gamma}_n - \gamma\right) = O_p(1)$, then*

$$\tilde{\mathcal{I}}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right) = \mathcal{I} + o_{p_{\delta,\gamma}}(1),$$

$$\sqrt{n}\left[\tilde{\delta}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right) - \delta\right] = \Delta_n^{\delta} + o_{p_{\delta,\gamma}}(1).$$

Proof of Theorem 1. Apply Theorem A.1(a) with $\delta = 0$ and $d_n = \mu(\beta, c)$. □

Table 2
Empirical size-$n = 1000$.

Bandwidth	$v_i \sim \mathcal{N}(0, 1) - \zeta = 1$			$v_i \sim \mathcal{N}(0, 1) - \zeta = 10$			$v_i \sim t(3) - \zeta = 1$			$v_i \sim t(3) - \zeta = 10$		
Scale (c)	AR	LM	LR	AR	LM	LR	AR	LM	LR	AR	LM	LR
0.45	0.161	0.122	0.153	0.153	0.097	0.113	0.078	0.060	0.071	0.075	0.063	0.070
0.46	0.150	0.115	0.144	0.139	0.091	0.106	0.072	0.056	0.066	0.068	0.059	0.066
0.47	0.137	0.109	0.136	0.129	0.087	0.101	0.066	0.054	0.062	0.063	0.056	0.063
0.48	0.130	0.105	0.129	0.119	0.084	0.096	0.062	0.051	0.058	0.060	0.054	0.059
0.49	0.122	0.099	0.121	0.109	0.079	0.092	0.058	0.049	0.056	0.057	0.051	0.058
0.50	0.115	0.094	0.115	0.104	0.074	0.088	0.055	0.048	0.053	0.052	0.049	0.056
0.51	0.106	0.090	0.106	0.098	0.072	0.083	0.052	0.046	0.050	0.050	0.048	0.053
0.52	0.098	0.085	0.101	0.092	0.069	0.080	0.048	0.044	0.046	0.048	0.046	0.052
0.53	0.092	0.081	0.093	0.085	0.065	0.078	0.044	0.043	0.044	0.045	0.045	0.049
0.54	0.087	0.078	0.090	0.079	0.062	0.073	0.042	0.041	0.042	0.043	0.043	0.048
0.55	0.081	0.074	0.085	0.073	0.060	0.069	0.040	0.039	0.040	0.040	0.040	0.045
0.56	0.077	0.072	0.078	0.068	0.057	0.066	0.037	0.039	0.038	0.038	0.039	0.043
0.57	0.072	0.068	0.076	0.064	0.054	0.063	0.034	0.038	0.037	0.036	0.037	0.042
0.58	0.068	0.065	0.072	0.060	0.051	0.061	0.032	0.036	0.035	0.034	0.036	0.039
0.59	0.065	0.064	0.071	0.056	0.050	0.058	0.030	0.035	0.034	0.033	0.035	0.038
0.60	0.060	0.062	0.067	0.053	0.049	0.057	0.028	0.034	0.032	0.031	0.035	0.037
0.61	0.058	0.059	0.063	0.050	0.046	0.055	0.026	0.033	0.030	0.028	0.034	0.037
0.62	0.054	0.057	0.060	0.047	0.044	0.053	0.024	0.031	0.029	0.027	0.033	0.035
0.63	0.051	0.055	0.057	0.044	0.043	0.052	0.022	0.031	0.027	0.026	0.032	0.035
0.64	0.047	0.053	0.054	0.042	0.042	0.050	0.021	0.030	0.027	0.024	0.031	0.034
0.65	0.044	0.051	0.051	0.041	0.041	0.048	0.020	0.029	0.025	0.023	0.030	0.033
0.66	0.042	0.050	0.049	0.038	0.039	0.046	0.019	0.028	0.023	0.022	0.029	0.032
0.67	0.040	0.047	0.047	0.037	0.036	0.044	0.018	0.027	0.022	0.021	0.029	0.030
0.68	0.038	0.046	0.045	0.035	0.035	0.043	0.016	0.025	0.021	0.021	0.029	0.030
0.69	0.037	0.045	0.042	0.034	0.035	0.041	0.015	0.024	0.020	0.020	0.029	0.030
0.70	0.035	0.044	0.040	0.031	0.033	0.039	0.014	0.024	0.019	0.018	0.028	0.029
0.71	0.033	0.043	0.038	0.029	0.032	0.037	0.013	0.023	0.018	0.017	0.028	0.029
0.72	0.032	0.042	0.035	0.029	0.031	0.036	0.013	0.023	0.017	0.016	0.027	0.029
0.73	0.029	0.040	0.033	0.028	0.031	0.035	0.012	0.023	0.016	0.016	0.026	0.028
0.74	0.028	0.039	0.031	0.027	0.030	0.034	0.011	0.022	0.015	0.015	0.025	0.028
0.75	0.028	0.037	0.030	0.025	0.029	0.033	0.011	0.022	0.014	0.015	0.024	0.027
0.76	0.027	0.036	0.030	0.024	0.029	0.033	0.011	0.021	0.013	0.015	0.024	0.026
0.77	0.025	0.035	0.029	0.022	0.028	0.032	0.010	0.021	0.013	0.014	0.024	0.026
0.78	0.024	0.035	0.027	0.021	0.027	0.031	0.010	0.021	0.012	0.012	0.024	0.026
0.79	0.023	0.034	0.026	0.019	0.026	0.030	0.010	0.021	0.012	0.012	0.023	0.025
0.80	0.022	0.034	0.025	0.018	0.026	0.030	0.009	0.021	0.011	0.012	0.022	0.024

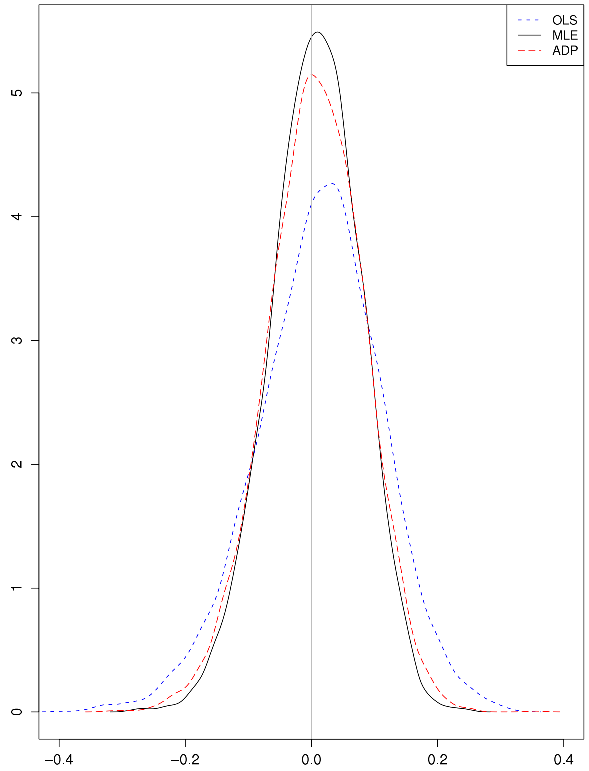

Fig. 4. Estimators of β.

Proof of Theorems 2, 4, and 6. Theorems 2 and 4(a) are special cases of Theorem 6(a) and 4(b) is a special case of Theorem 6(b), so it suffices to prove Theorem 6. In turn, Theorem 6 can be derived with the help of Theorem A.2 because $\hat{\Delta}_n^{**} = \sqrt{n}\tilde{\delta}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right)$ and $\hat{\mathcal{I}}_n^{**} = \tilde{\mathcal{I}}_n\left(\hat{\delta}_n, \hat{\gamma}_n\right)$, where $\hat{\delta}_n = \left(\hat{\Pi}_n', \hat{\pi}_n'\right)'$ and $\hat{\gamma}_n$ is as in the main text.

Proof of Theorems 6(a). If $c = 0$ in Assumption 4W, then the result can be obtained by applying Theorem A.2 with $\delta = \left(0', 0'\right)'$. The result for $c \neq 0$ follows by the contiguity property implied by Theorem A.1(a).

Proof of Theorems 6(b). If $b = 0$ in Assumption 4SC, then the result can be obtained by applying Theorem A.2 with $\delta = \left(0', \pi'\right)'$. The result for $b \neq 0$ follows by applying Theorem A.1(a) with $d_n = \left(b\pi', 0'\right)'$ and using Le Cam's third lemma.

Proof of Theorems 6(c). Apply Theorem A.2 with $\delta = (\beta\pi', \pi')'$. □

Proof of Theorems A.1. For every $\theta \in \mathbb{R}^2$, let $\bar{R}(\theta) = \theta'\mathcal{I}\theta/4 + \int_{\mathbb{R}^2} R(v, \theta) f(v)\, dv$, where, for $v \in \mathbb{R}^2$, $R(v, \theta) = 2\left[\sqrt{f(v-\theta)/f(v)} - 1 - \frac{1}{2}\theta'\ell(v)\right] 1\left[f(v) > 0\right]$. □

If Assumption 2 holds, then

$$\sqrt{f(v-\theta)} - \sqrt{f(v)} = \frac{1}{2}\theta' \int_0^1 \ell(v - \theta t)\sqrt{f(v - \theta t)}dt$$

$\forall v, \theta \in \mathbb{R}^2$,

and for almost every $v \in \mathbb{R}^2$, $\sqrt{f}$ is differentiable at v, with total derivative $-\frac{1}{2}\ell\sqrt{f}$. Using these facts and proceeding as in the

Table 3
Empirical size-$n = 5000$.

Bandwidth	$v_i \sim \mathcal{N}(0,1) - \zeta = 1$			$v_i \sim \mathcal{N}(0,1) - \zeta = 10$			$v_i \sim t(3) - \zeta = 1$			$v_i \sim t(3) - \zeta = 10$		
Scale (c)	AR	LM	LR	AR	LM	LR	AR	LM	LR	AR	LM	LR
0.45	0.126	0.098	0.127	0.139	0.091	0.106	0.064	0.059	0.066	0.071	0.063	0.068
0.46	0.118	0.095	0.117	0.128	0.088	0.102	0.059	0.056	0.062	0.066	0.062	0.066
0.47	0.111	0.091	0.109	0.120	0.083	0.096	0.055	0.054	0.058	0.061	0.059	0.063
0.48	0.100	0.086	0.102	0.111	0.081	0.093	0.051	0.051	0.055	0.056	0.057	0.061
0.49	0.093	0.080	0.096	0.106	0.077	0.089	0.049	0.048	0.052	0.053	0.054	0.059
0.50	0.086	0.077	0.088	0.100	0.075	0.085	0.046	0.047	0.047	0.050	0.052	0.057
0.51	0.081	0.075	0.081	0.093	0.073	0.084	0.044	0.045	0.045	0.048	0.049	0.055
0.52	0.074	0.072	0.076	0.089	0.071	0.081	0.041	0.043	0.043	0.045	0.048	0.053
0.53	0.069	0.069	0.072	0.084	0.070	0.078	0.039	0.042	0.041	0.042	0.046	0.052
0.54	0.064	0.066	0.068	0.078	0.067	0.077	0.037	0.041	0.039	0.040	0.045	0.050
0.55	0.060	0.064	0.063	0.072	0.064	0.075	0.036	0.039	0.038	0.038	0.045	0.048
0.56	0.057	0.061	0.060	0.067	0.063	0.073	0.033	0.038	0.036	0.037	0.043	0.048
0.57	0.051	0.059	0.058	0.063	0.060	0.069	0.032	0.037	0.034	0.035	0.041	0.046
0.58	0.048	0.058	0.054	0.060	0.056	0.067	0.030	0.036	0.032	0.033	0.040	0.045
0.59	0.046	0.056	0.052	0.058	0.055	0.064	0.028	0.035	0.031	0.033	0.037	0.043
0.60	0.044	0.054	0.050	0.055	0.053	0.062	0.027	0.034	0.029	0.031	0.036	0.041
0.61	0.043	0.052	0.048	0.053	0.052	0.060	0.026	0.032	0.028	0.030	0.034	0.039
0.62	0.040	0.050	0.045	0.050	0.051	0.058	0.025	0.032	0.027	0.029	0.034	0.037
0.63	0.038	0.048	0.044	0.048	0.049	0.055	0.024	0.031	0.026	0.028	0.033	0.036
0.64	0.036	0.047	0.042	0.045	0.048	0.054	0.023	0.030	0.025	0.027	0.031	0.036
0.65	0.036	0.045	0.041	0.044	0.047	0.053	0.022	0.030	0.024	0.025	0.031	0.035
0.66	0.034	0.043	0.039	0.042	0.046	0.052	0.022	0.029	0.023	0.024	0.030	0.034
0.67	0.033	0.042	0.038	0.040	0.045	0.050	0.021	0.027	0.023	0.023	0.030	0.032
0.68	0.032	0.041	0.036	0.038	0.043	0.049	0.021	0.026	0.023	0.022	0.029	0.031
0.69	0.031	0.040	0.034	0.037	0.042	0.048	0.019	0.026	0.021	0.021	0.028	0.030
0.70	0.029	0.039	0.033	0.036	0.041	0.047	0.018	0.026	0.021	0.021	0.027	0.030
0.71	0.027	0.038	0.032	0.035	0.041	0.046	0.018	0.025	0.020	0.020	0.027	0.029
0.72	0.026	0.038	0.031	0.033	0.039	0.046	0.018	0.024	0.019	0.019	0.026	0.028
0.73	0.025	0.037	0.031	0.032	0.039	0.044	0.017	0.023	0.019	0.018	0.025	0.027
0.74	0.024	0.036	0.030	0.031	0.038	0.044	0.017	0.024	0.019	0.017	0.024	0.027
0.75	0.022	0.035	0.029	0.029	0.037	0.042	0.016	0.023	0.019	0.017	0.024	0.027
0.76	0.022	0.034	0.028	0.028	0.036	0.041	0.016	0.023	0.018	0.016	0.023	0.026
0.77	0.021	0.033	0.027	0.026	0.035	0.040	0.015	0.022	0.017	0.015	0.022	0.025
0.78	0.020	0.032	0.025	0.025	0.035	0.040	0.014	0.022	0.017	0.014	0.022	0.025
0.79	0.020	0.031	0.024	0.025	0.035	0.039	0.014	0.021	0.017	0.014	0.021	0.024
0.80	0.019	0.030	0.023	0.024	0.034	0.038	0.012	0.021	0.016	0.013	0.021	0.023

proof of van der Vaart (1998, Lemma 7.6), it can be shown that if Assumption 2 holds, then

$$\lim_{\eta \downarrow 0} V(\eta) = 0,$$

$$V(\eta) = \sup_{\|\theta\| \leq \eta, \theta \neq 0} \|\theta\|^{-2} \int_{\mathbb{R}^2} R(v,\theta)^2 f(v)\, dv. \tag{5}$$

It follows from this result and Lemma 1 of Pollard (1997) that

$$\lim_{\eta \downarrow 0} \bar{V}(\eta) = 0, \quad \bar{V}(\eta) = \sup_{\|\theta\| \leq \eta, \theta \neq 0} \|\theta\|^{-2} \bar{R}(\theta). \tag{6}$$

The proofs of parts (a) and (b) are completely analogous, so to conserve space we only establish part (a). The log likelihood ratio $\mathcal{L}_n(d_n, 0)$ admits the expansion

$$\mathcal{L}_n(d_n,0) = d_n'(\mathcal{I} \otimes Q_{zz})\Delta_n^{\delta} + \sum_{i=1}^{n} R_{i,n} - \frac{1}{4}\sum_{i=1}^{n}\left[d_n'\frac{\ell_i \otimes z_i}{\sqrt{n}} + R_{i,n}\right]^2 \left(1+\xi_{i,n}\right),$$

where

$$R_{i,n} = R\left[\begin{pmatrix} y_{1i} - \gamma_1' x_i - \delta_1' z_i \\ y_{2i} - \gamma_2' x_i - \delta_2' z_i \end{pmatrix}, \begin{pmatrix} d_{1n}' z_i/\sqrt{n} \\ d_{2n}' z_i/\sqrt{n} \end{pmatrix}\right],$$

$$\xi_{i,n} = \xi\left[d_n'\frac{\ell_i \otimes z_i}{\sqrt{n}} + R_{i,n}\right],$$

and the defining property of $\xi(\cdot)$ is $\log(1+t) = t - \frac{1}{2} t^2 [1 + \xi(2t)]$.

It suffices to show that the following conditions hold:

$$\sum_{i=1}^{n} R_{i,n} = -\frac{1}{4} d_n'(\mathcal{I} \otimes Q_{zz}) d_n + o_{p_{\delta,\gamma}}(1), \tag{7}$$

$$\max_{1 \leq i \leq n} \left|\xi_{i,n}\right| = o_{p_{\delta,\gamma}}(1), \tag{8}$$

$$\sum_{i=1}^{n}\left[d_n'\frac{\ell_i \otimes z_i}{\sqrt{n}} + R_{i,n}\right]^2 = d_n'(\mathcal{I} \otimes Q_{zz}) d_n + o_{p_{\delta,\gamma}}(1). \tag{9}$$

To do so, suppose $(d, g) = (0, 0)$.

Proof of (7). The random variables $R_{1,n}, \ldots, R_{n,n}$ are independent and satisfy

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{E}\left(R_{i,n}^2\right) &\leq \sum_{i=1}^{n} V\left(\left\|\begin{pmatrix} d_{1n}' z_i/\sqrt{n} \\ d_{2n}' z_i/\sqrt{n} \end{pmatrix}\right\|^2\right)\left\|\begin{pmatrix} d_{1n}' z_i/\sqrt{n} \\ d_{2n}' z_i/\sqrt{n} \end{pmatrix}\right\|^2 \\ &\leq \max_{1 \leq i \leq n} V\left(\left\|\begin{pmatrix} d_{1n}' z_i/\sqrt{n} \\ d_{2n}' z_i/\sqrt{n} \end{pmatrix}\right\|^2\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} d_{1n}' z_i \\ d_{2n}' z_i \end{pmatrix}\right\|^2 \\ &= o(1)\, O(1) = o(1), \end{aligned}$$

where the penultimate equality uses (5) and Assumption 1(a). As a consequence,

$$\begin{aligned} \sum_{i=1}^{n} R_{i,n} &= \sum_{i=1}^{n} \mathbb{E}\left(R_{i,n}\right) + o_p(1) = -\frac{1}{4} d_n'\left(\mathcal{I} \otimes Q_{zz,n}\right) d_n \\ &\quad + \sum_{i=1}^{n} \bar{R}\left[\begin{pmatrix} d_{1n}' z_i/\sqrt{n} \\ d_{2n}' z_i/\sqrt{n} \end{pmatrix}\right] + o_p(1), \end{aligned}$$

where $d_n'\left(\mathcal{I}\otimes Q_{zz,n}\right)d_n = d_n'\left(\mathcal{I}\otimes Q_{zz}\right)d_n + o\left(1\right)$ by Assumption 1(a) and

$$\left|\sum_{i=1}^{n}\bar{R}\left[\begin{pmatrix} d_{1n}'z_i/\sqrt{n} \\ d_{1n}'z_i/\sqrt{n}\end{pmatrix}\right]\right| \leq \sum_{i=1}^{n}\left|\bar{R}\left[\begin{pmatrix} d_{1n}'z_i/\sqrt{n} \\ d_{2n}'z_i/\sqrt{n}\end{pmatrix}\right]\right|$$
$$\leq \sum_{i=1}^{n}\bar{V}\left(\left\|\begin{pmatrix} d_{1n}'z_i/\sqrt{n} \\ d_{2n}'z_i/\sqrt{n}\end{pmatrix}\right\|^2\right)\left\|\begin{pmatrix} d_{1n}'z_i/\sqrt{n} \\ d_{2n}'z_i/\sqrt{n}\end{pmatrix}\right\|^2$$
$$\leq \max_{1\leq i\leq n}\bar{V}\left(\left\|\begin{pmatrix} d_{1n}'z_i/\sqrt{n} \\ d_{2n}'z_i/\sqrt{n}\end{pmatrix}\right\|^2\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} d_{1n}'z_i \\ d_{2n}'z_i\end{pmatrix}\right\|^2$$
$$= o\left(1\right)O\left(1\right) = o\left(1\right),$$

where the penultimate equality uses (6) and Assumption 1(a).

Proof of (8). Because $\lim_{t\to 0}\xi\left(t\right) = 0$ (by Taylor's Theorem), the result follows from the fact that $\max_{1\leq i\leq n}\left\|\ell_i\otimes z_i\right\|/\sqrt{n} = o_p\left(1\right)$ (by $\ell_i \sim$ *i.i.d.* $(0,\mathcal{I})$ and Assumption 1(a)) and $\max_{1\leq i\leq n}\left|R_{i,n}\right| \leq \sqrt{\sum_{i=1}^{n}R_{i,n}^2} = o_p\left(1\right)$, where the latter uses the relation $\mathbb{E}\left(\sum_{i=1}^{n}R_{i,n}^2\right) = o\left(1\right)$ established in the proof of (7).

Proof of (9). Because $\sum_{i=1}^{n}R_{i,n}^2 = o_p\left(1\right)$ and

$$\sum_{i=1}^{n}\left[d_n'\frac{\ell_i\otimes z_i}{\sqrt{n}}\right]^2 = d_n'\left(\frac{1}{n}\sum_{i=1}^{n}\ell_i\ell_i'\otimes z_iz_i'\right)d_n,$$

it suffices to show that $n^{-1}\sum_{i=1}^{n}\ell_i\ell_i'\otimes z_iz_i' = \mathcal{I}\otimes Q_{zz} + o_p\left(1\right)$. The latter result can be established using $\ell_i \sim$ *i.i.d.* $(0,\mathcal{I})$ and Assumption 1(a). □

Proof of Theorem A.2. The proof uses Schick's (1987) approach.

First, it follows from Theorem A.1(b) and the properties of $\left(\hat{\delta}_n,\hat{\gamma}_n\right)$ that we may assume $\left(\hat{\delta}_n,\hat{\gamma}_n\right) = \left(\delta,\gamma\right)$. (This is so because Theorem 6.2 of Bickel (1982) can be used to verify that Condition A of Schick's (1987) Method 3 holds). In other words, it suffices to show that

$$\check{\Delta}_n^{\delta} = \left[\check{\mathcal{I}}_n^{-1}\otimes Q_{zz,n}^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\check{\ell}_n\left(v_i\right)\otimes z_i = \Delta_n^{\delta} + o_p\left(1\right) \qquad (10)$$

and

$$\check{\mathcal{I}}_n = \frac{1}{n}\sum_{i=1}^{n}\check{\ell}_n\left(v_i\right)\check{\ell}_n\left(v_i\right)' = \mathcal{I} + o_p\left(1\right), \qquad (11)$$

where

$$\check{\ell}_n\left(v\right) = -\frac{\partial\check{f}_n\left(v\right)/\partial v}{\check{f}_n\left(v\right)+a_n}, \qquad \check{f}_n\left(v\right) = \frac{1}{nh_n^2}\sum_{i=1}^{n}K\left(\frac{v-v_i}{h_n}\right).$$

To do so, let $\check{\ell}_{n,i}\left(\cdot\right)$ denote the leave-one-out version of $\check{\ell}_n\left(\cdot\right)$ given by

$$\check{\ell}_{n,i}\left(v\right) = -\frac{\partial\check{f}_{n,i}\left(v\right)/\partial v}{\check{f}_{n,i}\left(v\right)+a_n},$$
$$\check{f}_{n,i}\left(v\right) = \check{f}_n\left(v\right) - \frac{1}{nh_n^2}\left[K\left(\frac{v-v_i}{h_n}\right) - K\left(0\right)\right].$$

It follows from (the proof of) Lemma 3.1 and Remark 3.2 of Schick (1987) that condition (10) is implied by condition (11), Assumption 1(a) and 2, and the following conditions:

$$\mathbb{E}\left[\int_{\mathbb{R}^2}\left\|\check{\ell}_n\left(v\right)-\ell\left(v\right)\right\|^2 f\left(v\right)dv\right] = o\left(1\right), \qquad (12)$$

$$\max_{1\leq i\leq n}\mathbb{E}\left[\int_{\mathbb{R}^2}\left\|\check{\ell}_n\left(v\right)-\check{\ell}_{n,i}\left(v\right)\right\|^2 f\left(v\right)dv\right] = o\left(n^{-1}\right). \qquad (13)$$

Utilizing Assumptions 2 and 5 and proceeding as in Schick (1987, p. 100), it can be shown that

$$\int_{\mathbb{R}^2}\left\|\ell_n\left(v\right)-\ell\left(v\right)\right\|^2 f\left(v\right)dv = o\left(1\right), \qquad (14)$$

where

$$\ell_n\left(v\right) = -\frac{\partial f_n\left(v\right)/\partial v}{f_n\left(v\right)+a_n},$$
$$f_n\left(v\right) = \int_{\mathbb{R}^2}f\left(v-h_nr\right)K\left(r\right)dr = \mathbb{E}\left[\check{f}_n\left(v\right)\right].$$

It follows from this result that

$$\int_{\mathbb{R}^2}\left\|\ell_n\left(v\right)\right\|^2 f\left(v\right)dv = O\left(1\right). \qquad (15)$$

Using Assumptions 2 and 5 we have, uniformly in $v \in \mathbb{R}^2$, $\mathbb{E}[\|\check{f}_n\left(v\right) - f_n\left(v\right)\|^2] = O\left(n^{-1}h_n^{-2}\right)$ and $\mathbb{E}[\|\partial\check{f}_n\left(v\right)/\partial v - \partial f_n\left(v\right)/\partial v\|^2] = O\left(n^{-1}h_n^{-4}\right)$. Utilizing these facts, (15), and the decomposition

$$\check{\ell}_n\left(v\right)-\ell_n\left(v\right)$$
$$= -\ell_n\left(v\right)\frac{\check{f}_n\left(v\right)-f_n\left(v\right)}{\check{f}_n\left(v\right)+a_n} - \frac{\partial\check{f}_n\left(v\right)/\partial v - \partial f_n\left(v\right)/\partial v}{\check{f}_n\left(v\right)+a_n},$$

it is easily shown that

$$\int_{\mathbb{R}^2}\mathbb{E}\left[\left\|\check{\ell}_n\left(v\right)-\ell_n\left(v\right)\right\|^2\right]f\left(v\right)dv$$
$$= O\left(n^{-1}a_n^{-2}h_n^{-4}\right) = o\left(1\right), \qquad (16)$$

a result which can be combined with (14) to yield (12).

It follows from (15) to (16) that $\int_{\mathbb{R}^2}\mathbb{E}\left[\left\|\check{\ell}_n\left(v\right)\right\|^2\right]f\left(v\right)dv = O\left(1\right)$. Utilizing this fact, Assumption 5, and the decomposition

$$\check{\ell}_n\left(v\right)-\check{\ell}_{n,i}\left(v\right)$$
$$= \check{\ell}_n\left(v\right)\frac{\check{f}_n\left(v\right)-\check{f}_{n,i}\left(v\right)}{\check{f}_{n,i}\left(v\right)+a_n} + \frac{\partial\check{f}_n\left(v\right)/\partial v - \partial\check{f}_{n,i}\left(v\right)/\partial v}{\check{f}_{n,i}\left(v\right)+a_n},$$

it is easily shown that (13) holds.

Finally, condition (11) holds because

$$\check{\mathcal{I}}_n = \frac{1}{n}\sum_{i=1}^{n}\check{\ell}_{n,i}\left(v_i\right)\check{\ell}_{n,i}\left(v_i\right)' = \frac{1}{n}\sum_{i=1}^{n}\ell_i\ell_i' + o_p\left(1\right)$$
$$= \mathcal{I} + o_p\left(1\right),$$

where the first equality uses the fact that $\check{\ell}_{n,i}\left(v_i\right) = \check{\ell}_n\left(v_i\right)$ for each i and the second equality uses (14) and (16). □

Remarks. (i) Conditions (11) and (13) are counterparts of Schick's (1987) conditions (3.2) and (3.6). No counterpart of Schick's (1987) condition (3.1) is needed because $n^{-1}\sum_{i=1}^{n}z_i \to 0$. Also, the present definition of $\check{\ell}_{n,i}$ ensures that $\check{\ell}_{n,i}\left(v_i\right) = \check{\ell}_n\left(v_i\right)$ for every i, implying in particular that the natural counterpart of Schick's (1987) condition (3.5) is satisfied.

(ii) With the possible exception of (14), all steps in the proof of Theorem A.2 remain valid if the condition $\sup_{r\in\mathbb{R}}\left|k'\left(r\right)\right|/k\left(r\right) < \infty$ of Assumption 5(a) is replaced by the condition $\int_{\mathbb{R}}k'\left(r\right)^2dr < \infty$. The latter condition, which is implied by Assumption 5(a), is satisfied by the normal kernel. Furthermore, if the error density f

is such that $\sup_{v\in\mathbb{R}^2}\left\|\dot{f}\left(v\right)\right\| < \infty$, then (14) is satisfied (for any kernel) provided $\overline{\lim}_{n\to\infty} h_n/a_n < \infty$. This is so because

$$\begin{aligned}
&\int_{\mathbb{R}^2}\left\|\ell_n\left(v\right)-\ell\left(v\right)\right\|^2 f\left(v\right)dv\\
&\quad\le 2\int_{\mathbb{S}_f}\left\|\ell_n\left(v\right)\right\|^2\left[\sqrt{f\left(v\right)}-\sqrt{f_n\left(v\right)}\right]^2 dv\\
&\qquad+2\int_{\mathbb{S}_f}\left\|\ell_n\left(v\right)\sqrt{f_n\left(v\right)}-\ell\left(v\right)\sqrt{f\left(v\right)}\right\|^2 dv\\
&\quad=\left(\sup_{v\in\mathbb{R}^2}\left\|\dot{f}\left(v\right)\right\|\right)^2 o\left(h_n^2/a_n^2\right)+o\left(1\right),
\end{aligned}$$

where $\mathbb{S}_f = \left\{v\in\mathbb{R}^2 : f\left(v\right) > 0\right\}$ and the last equality uses

$$\int_{\mathbb{S}_f}\left[\sqrt{f\left(v\right)}-\sqrt{f_n\left(v\right)}\right]^2 dv = o\left(h_n^2\right), \tag{17}$$

$$\int_{\mathbb{S}_f}\left\|\ell_n\left(v\right)\sqrt{f_n\left(v\right)}-\ell\left(v\right)\sqrt{f\left(v\right)}\right\|^2 dv = o\left(1\right), \tag{18}$$

and the bound $\sup_{v\in\mathbb{R}^2}\left\|\ell_n\left(v\right)\right\|^2 \le \left(\sup_{v\in\mathbb{R}^2}\left\|\dot{f}\left(v\right)\right\|\right)^2/a_n^2$. The result (17) can be shown by means of Proposition A.7 of Koul and Schick (1996), while (18) can be established using Vitali's theorem, the L^1-continuity theorem, and arguments analogous to those used in the proof of Lemma 6.2 of Bickel (1982).

References

Amemiya, T., 1982. Two stage least absolute deviations estimators. Econometrica 50, 689–711.

Anderson, T.W., Rubin, H., 1949. Estimation of the parameters of a single equation in a complete set of stochastic equations. Annals of Mathematical Statistics 20, 46–63.

Andrews, D.W.K., Marmer, V., 2007. Exactly distribution-free inference in instrumental variables regression with possibly weak instruments. Journal of Econometrics 142, 183–200.

Andrews, D.W.K., Moreira, M.J., Stock, J.H., 2006. Optimal two-sided invariant similar tests for instrumental variables regression. Econometrica 74, 715–752.

Andrews, D.W.K., Soares, G., 2007. Rank tests for instrumental variables regression with weak instruments. Econometric Theory 23, 1033–1082.

Andrews, D.W.K., Stock, J.H., 2007. Inference with weak instruments. In: Blundell, R., Newey, W.K., Persson., T. (Eds.), Advances in Economics and Econometrics: Theory and Applications. In: Ninth World Congress, vol. III. Cambridge University Press, New York, pp. 122–173.

Bickel, P.J., 1982. On adaptive estimation. Annals of Statistics 10, 647–671.

Choi, S., Hall, W.J., Schick, A., 1996. Asymptotically uniformly most powerful tests in parametric and semiparametric models. Annals of Statistics 24, 841–861.

Cox, D.R., Reid, N.R., 1987. Parameter orthogonality and approximate conditional inference with discussion. Journal of the Royal Statistical Society, Series B 49, 1–39.

Dufour, J.-M., 2003. Identification, Weak instruments, and statistical inference in econometrics. Canadian Journal of Economics 36, 767–808.

Fuller, W.A., 1977. Some properties of a modification of the limited information estimator. Econometrica 45, 939–953.

Ichimura, H., Todd, P.E., 2007. In: Heckman, J.J., Leamer., E.E. (Eds.), Implementing Nonparametric and Semiparametric Estimators. In: Handbook of Econometrics, vol. 6B. North Holland, New York, pp. 5369–5468.

Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. Econometrica 70, 1781–1803.

Koul, H.L., Schick, A., 1996. Adaptive estimation in a random coefficient autoregressive model. Annals of Statistics 24, 1025–1052.

Moreira, M.J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71, 1027–1048.

Pollard, D., 1997. Another look at differentiability in quadratic mean. In: Pollard, D., Torgersen, E., Yang., G.L. (Eds.), Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics. Springer-Verlag, New York, pp. 305–314.

Powell, J.L., 1983. The asymptotic normality of two-stage least absolute deviations estimators. Econometrica 51, 1569–1575.

Schick, A., 1987. A note on the construction of asymptotically linear estimators. Journal of Statistical Planning and Inference 16, 89–105.

Staiger, D., Stock, J.H., 1997. Instrumental variables estimation with weak instruments. Econometrica 65, 557–586.

Steigerwald, D.G., 1992. On the finite sample behavior of adaptive estimators. Journal of Econometrics 54, 371–400.

van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, New York.