

Optimal Inference for Instrumental Variables Regression with non-Gaussian Errors*

MATIAS D. CATTANEO
DEPARTMENT OF ECONOMICS
UC BERKELEY

RICHARD K. CRUMP
DEPARTMENT OF ECONOMICS
UC BERKELEY

MICHAEL JANSSON
DEPARTMENT OF ECONOMICS
UC BERKELEY

PRELIMINARY AND INCOMPLETE DRAFT

November 2006

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 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.^{1,2} Analyzing such a model in which the *i.i.d.* errors are

*The authors thank Jim Powell, Paul Ruud, Hal White, and seminar participants at Michigan and UCSD for comments. The third author gratefully acknowledges the research support of Danish Social Sciences Research Council and the hospitality of the University of Aarhus.

¹Important contributions to the weak IV literature include Nelson and Startz (1990a, 1990b), Bound, Jaeger, and Baker (1995), Dufour (1997), Staiger and Stock (1997), Kleibergen (2002), Moreira (2003), and Andrews, Moreira, and Stock (2006a). For reviews, see e.g. Stock, Wright, and Yogo (2002), Dufour (2003), Hahn and Hausman (2003), and Andrews and Stock (2006a).

²Improved understanding of the properties of models with a “large” number of instruments has also been achieved in recent years. Papers on that topic include Bekker (1994), Hahn (2002), Chao

furthermore assumed to be Gaussian, Andrews, Moreira, and Stock (2006a) 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 Andrews, Moreira, and Stock (2006a) (and others). Recent work by Andrews and Marmar (2005) and Andrews and Soares (2006) 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 be decomposed into two separate problems, each of which is well understood (in isolation) from the works of Bickel (1982) and Andrews, Moreira, and Stock (2006a), 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

and Swanson (2005), Stock and Yogo (2005), Hansen, Hausman, and Newey (2005), Andrews and Stock (2006b), and Chioda and Jansson (2006). The present paper does not employ many (weak) instruments asymptotics, but it would be of interest to generalize our results along those lines.

Andrews, Moreira, and Stock (2006a) to construct test statistics which combine the efficient estimators of the reduced form regression coefficients in a “nearly” 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 interval 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 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. Mathematical derivations have been relegated to an Appendix.

2. THE MODEL

We consider a model given by

$$\begin{aligned} y_{1i} &= \Gamma_1' x_i + \beta y_{2i} + u_i, \\ y_{2i} &= \gamma_2' x_i + \pi' z_i + v_{2i} \quad (i = 1, \dots, n), \end{aligned} \tag{1}$$

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 $\Gamma_1, \gamma_2 \in \mathbb{R}^p$ are parameters. The exogenous variables x_i and z_i are fixed (i.e., nonrandom) and the first element of x_i is assumed to equal 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

$$\begin{aligned} y_{1i} &= \gamma'_1 x_i + \beta \pi' z_i + v_{1i}, \\ y_{2i} &= \gamma'_2 x_i + \pi' z_i + v_{2i} \quad (i = 1, \dots, n), \end{aligned} \quad (2)$$

where $\gamma_1 = \Gamma_1 + \gamma_2 \beta$ and $v_{1i} = v_{2i} \beta + u_i$. The parameters of the reduced form are β , π , $\gamma = (\gamma'_1, \gamma'_2)'$, 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

$$\begin{aligned} y_{1i} &= \gamma'_1 x_i + \delta'_1 z_i + v_{1i}, \\ y_{2i} &= \gamma'_2 x_i + \delta'_2 z_i + v_{2i} \quad (i = 1, \dots, n), \end{aligned} \quad (3)$$

where $\delta_1, \delta_2 \in \mathbb{R}^q$ and the other parameters are as in (2). (The model (3) reduces to (2) when $\delta = (\delta_1, \delta_2)' = (\beta \pi', \pi)'$.) Indeed, the main results of this paper can and will be derived as relatively simple consequences of results concerning the model (3), which itself can be analyzed by means of fairly standard tools.

Our goal is to develop powerful tests of

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

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

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

$$f(v + \theta) - f(v) = \theta' \int_0^1 \dot{f}(v + \theta t) dt, \quad \forall \theta \in \mathbb{R}^2.$$

³Testing 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.

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

(c) $\int_{\mathbb{R}^2} \ell(v)' \ell(v) f(v) dv < \infty$, where

$$\ell(v) = -\frac{\dot{f}(v)}{f(v)} 1[f(v) > 0].$$

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

Assumption 1 is a standard assumption concerning the exogenous variables. It holds in probability if the $(x_i', z_i)'$ are a realization of an *i.i.d.* sequence with positive definite variance matrix and finite second moment.⁵ 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.⁶ An immediate implication of Assumptions 1(a) and 2 is that

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

where

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

is the Fisher information for the location family generated by f .

Assumption 2 furthermore implies that the model (3) is differentiable in quadratic mean at any (γ, δ) (see (32) in the proof of Theorem A.1 in the Appendix) and enables nonparametric estimation of ℓ (as demonstrated by Theorem A.2 in the Appendix). In other words, the roles played by parts (a) and (b) of Assumptions 2 are analogous to those played by the assumption of absolute continuity routinely invoked in regression models with scalar errors. In fact, the scalar counterpart of Assumption 2(b) is the assumption of absolute continuity.⁷ In models where dependence between the errors v_{1i} and v_{2i} is allowed, the present smoothness condition imposed on the density would

⁵If the the exogenous variables $(x_i', z_i)'$ are treated as random, then independence of $\{v_i\}$ and $\{(x_i', z_i)'\}$ would be required for the results of this paper to remain valid. It seems plausible that certain forms of heteroskedasticity can be accommodated by adapting the methods of Schick (1997), but no attempts to do so will be made in this paper.

⁶Assumption 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.

⁷In the scalar context, but apparently not otherwise, it follows from Lebesgue's differentiation theorem (e.g., Dudley (2002, Theorem 7.2.1)) that the counterpart of Assumption 2(a) is implied by the counterpart Assumption 2(b).

appear to be more natural than the assumption of absolute continuity, the reason being that parts (a) and (b) of Assumption 2 involve only partial derivatives of order one, whereas the assumption of absolute continuity involves second-order cross partials.

Assumption 3 is a normalization which greatly simplifies the derivation and statements of asymptotic results. Specifically, because the limit of $Q_{xz,n}$ 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 $\sum_{i=1}^n z_i = 0$ 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 $\beta = \beta_0 + O(1/\sqrt{n})$. 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.⁸

⁸In addition, Assumption 4SF would be natural in investigations of the large deviation (efficiency) properties of inference procedures (e.g., Puhalskii and Spokoiny (1998)). We do not employ large deviations techniques in this paper.

Assumptions 4W, 4SC, and 4SF are nonnested, but it seems natural to study them in the order indicated above. This is so because the assumptions impose decreasingly strong restrictions on the parameters δ_1 and δ_2 of (3). Specifically, Assumption 4W implies that $\delta_1 = O(1/\sqrt{n})$ and $\delta_2 = O(1/\sqrt{n})$. Relative to Assumption 4W, Assumption 4SC removes the requirement $\delta_2 = O(1/\sqrt{n})$ and Assumption 4SF furthermore relaxes the requirement $\delta_1 = O(1/\sqrt{n})$. 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 Assumption 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

$$\Omega = \int_{\mathbb{R}^2} vv' f(v) dv \quad (5)$$

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 Andrews, Moreira, and Stock (2006a)). This feature is lost, in general, when f is not Gaussian. On the other hand, 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 $\mathcal{N}[\mu(b, c), \Omega \otimes Q_{zz}^{-1}]$ distribution, where

$$\mu(b, c) = \begin{pmatrix} b \\ 1 \end{pmatrix} \otimes c. \quad (6)$$

Because $\Omega = \mathcal{I}^{-1}$ when f is Gaussian, an equivalent characterization of the limiting experiment in the Gaussian case is that it is that of a single observation from the $\mathcal{N}[\mu(b, c), \mathcal{I}^{-1} \otimes Q_{zz}^{-1}]$ distribution. The latter characterization generalizes readily to

non-Gaussian error distributions. To give a precise statement, define the log likelihood ratio function

$$\begin{aligned}
 L_n(\beta, c) &= \sum_{i=1}^n \log f(y_{1i} - \gamma'_1 x_i - \beta c' z_i / \sqrt{n}, y_{2i} - \gamma'_2 x_i - c' z_i / \sqrt{n}) \\
 &\quad - \sum_{i=1}^n \log f(y_{1i} - \gamma'_1 x_i, y_{2i} - \gamma'_2 x_i),
 \end{aligned} \tag{7}$$

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

Theorem 1. *If Assumptions 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

$$\Delta_n = (\mathcal{I}^{-1} \otimes Q_{zz}^{-1}) \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell(y_{1i} - \gamma'_1 x_i, y_{2i} - \gamma'_2 x_i) \otimes z_i \rightarrow_{d_0} \mathcal{N}(0, \mathcal{I}^{-1} \otimes Q_{zz}^{-1}).$$

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.

Theorem 1 and Le Cam’s third lemma can be used to show that if Assumptions 1(a), 2, and 4W hold, then the asymptotically sufficient statistic Δ_n satisfies

$$\Delta_n \rightarrow_d \mathcal{N}[\mu(\beta, c), \mathcal{I}^{-1} \otimes Q_{zz}^{-1}]. \tag{8}$$

In other words, the limiting experiment is that of a single observation from the $\mathcal{N}[\mu(\beta, c), \mathcal{I}^{-1} \otimes Q_{zz}^{-1}]$ distribution whether or not the errors are Gaussian.

Under the same assumptions, the (quasi-)sufficient statistic⁹

$$\bar{\Delta}_n = (\Omega \otimes Q_{zz,n}^{-1}) \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{\ell}(y_{1i} - \gamma'_1 x_i, y_{2i} - \gamma'_2 x_i) \otimes z_i, \quad \bar{\ell}(v) = \Omega^{-1} v,$$

⁹The statistic $\bar{\Delta}_n$ does not depend on Ω , but the present formulation facilitates comparison of $\bar{\Delta}_n$ and Δ_n and we therefore prefer it to one which does not involve Ω .

obtained from the Gaussian (quasi-)likelihood satisfies

$$\bar{\Delta}_n \rightarrow_d \mathcal{N} [\mu(\beta, c), \Omega \otimes Q_{zz}^{-1}].$$

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 (under weak identification and whenever the inequality $\mathcal{I}^{-1} \leq \Omega$ is strict) by a test which is efficient (or “nearly” so) under the assumptions of Theorem 1.¹⁰

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

$$\begin{pmatrix} S_n \\ T_n \end{pmatrix} = [\mathcal{I}^{1/2'} \otimes Q_{zz}^{1/2'}] \Delta_n, \quad (9)$$

where $M^{1/2}$ denotes the upper triangular Cholesky factor of a (symmetric, positive semi-definite) matrix M ; that is, $M = M^{1/2} M^{1/2'}$, where $M^{1/2}$ is upper triangular.¹¹ The pair (S_n, T_n) is a non-Gaussian counterpart of

$$\begin{pmatrix} \bar{S}_n \\ \bar{T}_n \end{pmatrix} = [(\Omega^{-1})^{1/2'} \otimes Q_{zz,n}^{1/2'}] \bar{\Delta}_n,$$

which features prominently in the work by Moreira (2003), Andrews, Moreira, and Stock (2006a), and others.

¹⁰Section 4 will exhibit tests which are “nearly” efficient under the assumptions of Theorem 1.

¹¹In particular, letting \mathcal{I}_{ij} denote element (i, j) of \mathcal{I} , we have:

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

In terms of (\bar{S}_n, \bar{T}_n) , 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

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

In perfect analogy with the Gaussian case, let

$$AR_n = S'_n S_n, \quad LM_n = \frac{(S'_n T_n)^2}{T'_n T_n}, \quad (10)$$

$$LR_n = \frac{1}{2} \left(S'_n S_n - T'_n T_n + \sqrt{(S'_n S_n - T'_n T_n)^2 + 4 (S'_n T_n)^2} \right). \quad (11)$$

The tests which reject H_0 when $AR_n > \chi^2_\alpha(q)$, $LM_n > \chi^2_\alpha(1)$, and $LR_n > \kappa_\alpha(T_n)$ have asymptotic size α , where $\chi^2_\alpha(d)$ 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 $\frac{1}{2} \left(\mathcal{Z}' \mathcal{Z} - t't + \sqrt{(\mathcal{Z}' \mathcal{Z} - t't)^2 + 4 (\mathcal{Z}' t)^2} \right)$, where $\mathcal{Z} \sim \mathcal{N}(0, I_q)$.¹² 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 Andrews, Moreira, and Stock (2006a). In particular, it follows from Andrews, Moreira, and Stock (2006a) 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.

Remark. The existence of tests which are equivalent to procedures based on the Gaussian quasi-likelihood under the assumption of normality and enjoy improved power properties for certain non-Gaussian error distributions has been pointed out by Andrews and Marmer (2005) and Andrews and Soares (2006). The rank-based tests proposed in those papers, while superior to tests based on the Gaussian quasi-likelihood under some conditions, are also inefficient in general (even for those error distributions for which they dominate tests based on the Gaussian quasi-likelihood).

¹²As shown by Moreira (2003), $\kappa_\alpha(t)$ depends on t only through $\|t\|$, is monotonically decreasing in $\|t\|$, and satisfies $\lim_{\|t\| \rightarrow \infty} \kappa_\alpha(t) = \chi^2_\alpha(1)$. The latter result will be utilized when studying the behavior of the test based on LR_n under strong identification.

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 (i.e., that π is “close” to zero). 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 $(\hat{\Delta}_n, \hat{\mathcal{I}}_n)$ which is asymptotically equivalent to (Δ_n, \mathcal{I}) under weak identification and can be computed without knowledge of (γ, f) . To that end, let

$$\hat{\Delta}_n = \left(\hat{\mathcal{I}}_n^{-1} \otimes Q_{zz,n}^{-1} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_{i,n} \otimes z_i, \quad \hat{\mathcal{I}}_n = n^{-1} \sum_{i=1}^n \hat{\ell}_{i,n} \hat{\ell}'_{i,n}, \quad (12)$$

where $\hat{\ell}_{i,n}$ is an estimator of $\ell(y_{1i} - \gamma'_1 x_i, y_{2i} - \gamma'_2 x_i)$. In the spirit of Schick (1987, 1993), we assume that $\hat{\ell}_{i,n} = \hat{\ell}_n(\hat{v}_i)$, where $\hat{v}_i = (y_{1i} - \hat{\gamma}'_{1n} x_i, y_{2i} - \hat{\gamma}'_{2n} x_i)'$ for some estimator $\hat{\gamma}_n = (\hat{\gamma}'_{1n}, \hat{\gamma}'_{2n})'$ of γ and

$$\hat{\ell}_n(v) = -\frac{\partial \hat{f}_n(v) / \partial v}{\hat{f}_n(v) + a_n}, \quad \hat{f}_n(v) = \frac{1}{nh_n^2} \sum_{i=1}^n K\left(\frac{v - \hat{v}_i}{h_n}\right), \quad (13)$$

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.^{13,14}

Assumption 5. (a) $K(s_1, s_2) = k(s_1)k(s_2)$, where k is a bounded, symmetric, continuously differentiable density satisfying

$$\int_{\mathbb{R}} r^2 k(r) dr < \infty \quad \text{and} \quad \sup_{r \in \mathbb{R}} |k'(r)| / k(r) < \infty.$$

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

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

¹³If 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.

¹⁴In Assumption 6, the statement “ $\hat{\gamma}_n$ is discrete” is shorthand for the assumption that $\hat{\gamma}_n$ takes only values in the grid $\{\varkappa Z / \sqrt{n} : Z \in \mathbb{Z}^{2p}\}$, where \varkappa is some (constant) positive definite diagonal $2p \times 2p$ matrix. A similar remark applies to Assumptions 7 and 8.

Assumption 5(a) holds if k is the logistic density, but not if k is the standard normal density, the reason being that the normal density violates the condition $\sup_{r \in \mathbb{R}} |k'(r)|/k(r) < \infty$. As explained in a remark 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 $\sup_{v \in \mathbb{R}^2} \|\dot{f}(v)\| < \infty$ (and provided the requirement $\overline{\lim}_{n \rightarrow \infty} h_n/a_n < \infty$ is added to Assumption 5(b)). Assumption 6 is satisfied (under both weak and strong identification) by a discretized version of

$$\hat{\gamma}_n^{OLS} = \begin{bmatrix} \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \left(\sum_{i=1}^n x_i y_{1i} \right) \\ \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \left(\sum_{i=1}^n x_i y_{2i} \right) \end{bmatrix},$$

the OLS estimator of γ .

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

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

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 (discrete) \sqrt{n} -consistent estimator of δ in that case.¹⁵ 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

$$\widehat{AR}_n = \hat{S}'_n \hat{S}_n, \quad \widehat{LM}_n = \frac{\left(\hat{S}'_n \hat{T}_n \right)^2}{\hat{T}'_n \hat{T}_n}, \quad (14)$$

and

$$\widehat{LR}_n = \frac{1}{2} \left(\hat{S}'_n \hat{S}_n - \hat{T}'_n \hat{T}_n + \sqrt{\left(\hat{S}'_n \hat{S}_n - \hat{T}'_n \hat{T}_n \right)^2 + 4 \left(\hat{S}'_n \hat{T}_n \right)^2} \right), \quad (15)$$

¹⁵The full force of Theorem A.2 will be needed when Assumption 4W is replaced by Assumption 4SC or 4SF.

respectively, where $(\hat{S}'_n, \hat{T}'_n)' = [\hat{\mathcal{T}}_n^{1/2'} \otimes Q_{zz,n}^{1/2'}] \hat{\Delta}_n$. More specifically, we have the following corollary, which implies in particular that the (feasible) test which rejects when $\widehat{LR}_n > \kappa_\alpha(\hat{T}_n)$ is “nearly efficient” when identification is weak.

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

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

Remark. The statistic $\hat{\Delta}_n/\sqrt{n}$ is a bivariate version of the adaptive estimator constructed by Schick (1987) (following the seminal work of Bickel (1982)). In scalar regression models, adaptive estimators of slope coefficients can alternatively be based on regression quantiles of Koenker and Bassett (1978) (Portnoy and Koenker (1989)), Hansen’s (1982) generalized method of moments (GMM) (Newey (1988, 2004)), or a series estimator of the log density function (Faraway (1992), Jin (1992)). It would be of potential interest to generalize these constructions to our bivariate setting.

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 (Δ_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 satisfies¹⁶

$$\bar{\Delta}_n - \sqrt{n} \begin{pmatrix} 0 \\ \pi \end{pmatrix} \rightarrow_d \mathcal{N} \left[\begin{pmatrix} b\pi \\ 0 \end{pmatrix}, \Omega \otimes Q_{zz}^{-1} \right].$$

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

$$\overline{AR}_n \rightarrow_d \chi^2(q; b^2 \omega_{11}^{-1} \pi' Q_{zz} \pi),$$

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

¹⁶The displayed property and the other asymptotic properties of $\bar{\Delta}_n$ mentioned in this paper are shared by feasible versions of $\bar{\Delta}_n$ which replace γ and Ω by estimators (provided the estimator of Ω is invertible). A similar remarks applies to the pair (\bar{S}_n, \bar{T}_n) and smooth functions thereof provided Ω is replaced by a consistent estimator.

and $\kappa_\alpha(\bar{T}_n) = \chi_\alpha^2(1) + o_p(1)$, 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 . 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 by¹⁷

$$\hat{\Delta}_n^* = \begin{pmatrix} 0 \\ \sqrt{n}\hat{\pi}_n \end{pmatrix} + \left(\hat{\mathcal{I}}_n^{*-1} \otimes Q_{zz,n}^{-1} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_{i,n}^* \otimes z_i, \quad \hat{\mathcal{I}}_n^* = n^{-1} \sum_{i=1}^n \hat{\ell}_{i,n}^* \hat{\ell}_{i,n}^{*'} \quad (16)$$

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

$$\hat{\ell}_n^*(v) = -\frac{\partial \hat{f}_n^*(v) / \partial v}{\hat{f}_n^*(v) + a_n}, \quad \hat{f}_n^*(v) = \frac{1}{nh_n^2} \sum_{i=1}^n K\left(\frac{v - \hat{v}_i^*}{h_n}\right). \quad (17)$$

As defined, $\hat{\Delta}_n^*/\sqrt{n}$ is a one-step estimator of δ (in the model (3)) which uses $(0', \hat{\pi}_n)'$ as an initial estimator of δ . This initial estimator is \sqrt{n} -consistent under Assumption 4SC provided $\hat{\pi}_n$ satisfies the following condition.

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

Assumption 7 holds (under both weak and strong identification) if $\hat{\pi}_n$ is a discretized version of $\hat{\pi}_n^{OLS} = (\sum_{i=1}^n z_i z_i')^{-1} (\sum_{i=1}^n z_i y_{2i})$.

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

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

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

$$\hat{\Delta}_n^* - \sqrt{n} \begin{pmatrix} 0 \\ \pi \end{pmatrix} \rightarrow_d \mathcal{N} \left[\begin{pmatrix} b\pi \\ 0 \end{pmatrix}, \mathcal{I}^{-1} \otimes Q_{zz}^{-1} \right].$$

¹⁷The pair $(\hat{\Delta}_n^*, \hat{\mathcal{I}}_n^*)$ reduces to $(\hat{\Delta}_n, \hat{\mathcal{I}}_n)$ when $\hat{\pi}_n = 0$. Moreover, $\sqrt{n}\pi = O(1)$ under weak identification, so Theorem 2 is a special case of Theorem 4 (a).

As a consequence of Theorem 4, we have the following result concerning the statistics

$$\widehat{AR}_n^* = \hat{S}_n^{*'} \hat{S}_n^*, \quad \widehat{LM}_n^* = \frac{(\hat{S}_n^{*'} \hat{T}_n^*)^2}{\hat{T}_n^{*'} \hat{T}_n^*}, \quad (18)$$

$$\widehat{LR}_n^* = \frac{1}{2} \left(\hat{S}_n^{*'} \hat{S}_n^* - \hat{T}_n^{*'} \hat{T}_n^* + \sqrt{(\hat{S}_n^{*'} \hat{S}_n^* - \hat{T}_n^{*'} \hat{T}_n^*)^2 + 4 (\hat{S}_n^{*'} \hat{T}_n^*)^2} \right), \quad (19)$$

where $(\hat{S}_n^{*'}, \hat{T}_n^{*'})' = [\hat{\mathcal{I}}_n^{*1/2'} \otimes Q_{zz,n}^{1/2'}] \hat{\Delta}_n^*$.

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

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

(b) *If Assumptions 1-3, 4SC, and 5-7 hold, then*

$$\begin{aligned} \widehat{AR}_n^* &= AR_n + o_p(1) \rightarrow_d \chi^2(q; b^2 \mathcal{I}_{11.2} \pi' Q_{zz} \pi), \\ \widehat{LM}_n^* &= \widehat{LR}_n^* + o_p(1) = \frac{(S_n' Q_{zz}^{1/2'} \pi)^2}{\pi' Q_{zz} \pi} + o_p(1) \rightarrow_d \chi^2(1; b^2 \mathcal{I}_{11.2} \pi' Q_{zz} \pi), \end{aligned}$$

and $\kappa_\alpha(\hat{T}_n^*) = \chi_\alpha^2(1) + o_p(1)$.

It follows from Corollary 5(a) that the test which rejects when $\widehat{LR}_n^* > \kappa_\alpha(\hat{T}_n^*)$ is “nearly efficient” when identification is weak. Moreover, Theorem A.1 in the Appendix and Choi, Hall, and Schick (1996, Theorem 2) can be used to show that the test which rejects for large values of $(S_n' Q_{zz}^{1/2'} \pi)^2 / (\pi' Q_{zz} \pi)$ is asymptotically uniformly most powerful unbiased (in the terminology of Choi, Hall, and Schick (1996, Section 4)) under the assumptions of Corollary 5(b).¹⁸ As a consequence, Corollary 5(b) implies that the test which rejects when $\widehat{LR}_n^* > \kappa_\alpha(\hat{T}_n^*)$ enjoys demonstrable

¹⁸As mentioned by Choi, Hall, and Schick (1996, p. 852), the test furthermore satisfies a location invariance property and is therefore also the asymptotically uniformly most powerful location invariant test. A potential advantage of imposing location invariance when defining optimality criteria is that the resulting notion of optimality is applicable in some nonstandard testing problems as well. In particular, it can be applied to derive attainable semiparametric power envelopes for the unit root testing problem (Jansson (2006)).

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's (2006) 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)

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

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

$$\bar{\Delta}_n - \sqrt{n} \begin{pmatrix} \beta\pi \\ \pi \end{pmatrix} \rightarrow_d \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega \otimes Q_{zz}^{-1} \right].$$

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, let¹⁹

$$\hat{\Delta}_n^{**} = \begin{pmatrix} \sqrt{n}\hat{\Pi}_n \\ \sqrt{n}\hat{\pi}_n \end{pmatrix} + \left(\hat{\mathcal{I}}_n^{**,-1} \otimes Q_{zz,n}^{-1} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_{i,n}^{**} \otimes z_i, \quad \hat{\mathcal{I}}_n^{**} = n^{-1} \sum_{i=1}^n \hat{\ell}_{i,n}^{**} \hat{\ell}_{i,n}^{**\prime} \quad (20)$$

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

$$\hat{\ell}_n^{**}(v) = -\frac{\partial \hat{f}_n^{**}(v) / \partial v}{\hat{f}_n^{**}(v) + a_n}, \quad \hat{f}_n^{**}(v) = \frac{1}{nh_n^2} \sum_{i=1}^n K \left(\frac{v - \hat{v}_i^{**}}{h_n} \right), \quad (21)$$

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

¹⁹The pair $(\hat{\Delta}_n^{**}, \hat{\mathcal{I}}_n^{**})$ reduces to $(\hat{\Delta}_n^*, \hat{\mathcal{I}}_n^*)$ when $\hat{\Pi}_n = 0$. Moreover, $\sqrt{n}\beta\pi = O(1)$ under the assumptions of Theorem 4, so that theorem is a special case of Theorem 6(a)-(b).

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*

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

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

$$\hat{\Delta}_n^{**} - \sqrt{n} \begin{pmatrix} 0 \\ \pi \end{pmatrix} \rightarrow_d \mathcal{N} \left[\begin{pmatrix} b\pi \\ 0 \end{pmatrix}, \mathcal{I}^{-1} \otimes Q_{zz}^{-1} \right].$$

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

$$\hat{\Delta}_n^{**} - \sqrt{n} \begin{pmatrix} \beta\pi \\ \pi \end{pmatrix} \rightarrow_d \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathcal{I}^{-1} \otimes Q_{zz}^{-1} \right].$$

Let $\left(\hat{S}_n^{**'}, \hat{T}_n^{**'} \right)' = \left[\hat{\mathcal{I}}_n^{**1/2'} \otimes Q_{zz,n}^{1/2'} \right] \hat{\Delta}_n^{**}$ and define

$$\widehat{AR}_n^{**} = \hat{S}_n^{**'} \hat{S}_n^{**}, \quad \widehat{LM}_n^{**} = \frac{\left(\hat{S}_n^{**'} \hat{T}_n^{**} \right)^2}{\hat{T}_n^{**'} \hat{T}_n^{**}}, \quad (22)$$

$$\widehat{LR}_n^{**} = \frac{1}{2} \left(\hat{S}_n^{**'} \hat{S}_n^{**} - \hat{T}_n^{**'} \hat{T}_n^{**} + \sqrt{\left(\hat{S}_n^{**'} \hat{S}_n^{**} - \hat{T}_n^{**'} \hat{T}_n^{**} \right)^2 + 4 \left(\hat{S}_n^{**'} \hat{T}_n^{**} \right)^2} \right). \quad (23)$$

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*

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

(b) *If Assumptions 1-3, 4SC, and 5-8 hold, then*

$$\widehat{AR}_n^{**} = AR_n + o_p(1) \rightarrow_d \chi^2 \left(q; b^2 \mathcal{I}_{11.2} \pi' Q_{zz} \pi \right),$$

$$\widehat{LM}_n^{**} = \widehat{LR}_n^{**} + o_p(1) = \frac{\left(S_n' Q_{zz}^{1/2'} \pi\right)^2}{\pi' Q_{zz} \pi} + o_p(1) \rightarrow_d \chi^2\left(1; b^2 \mathcal{I}_{11.2} \pi' Q_{zz} \pi\right),$$

and $\kappa_\alpha\left(\widehat{T}_n^{**}\right) = \chi_\alpha^2(1) + o_p(1)$.

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

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

In perfect analogy with Corollary 5, parts (a) and (b) of Corollary 7 imply that the test which rejects when $\widehat{LR}_n^{**} > \kappa_\alpha\left(\widehat{T}_n^{**}\right)$ 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 $\widehat{LR}_n^* > \kappa_\alpha\left(\widehat{T}_n^*\right)$, 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

$$\widehat{\beta}_n^{**} = \frac{\widehat{\Delta}_{1,n}^{**'} Q_{zz,n} \widehat{\Delta}_{2,n}^{**}}{\widehat{\Delta}_{2,n}^{**'} Q_{zz,n} \widehat{\Delta}_{2,n}^{**}}, \quad (24)$$

where $\widehat{\Delta}_n^{**} = \left(\widehat{\Delta}_{1,n}^{**'}, \widehat{\Delta}_{2,n}^{**'}\right)'$ and partitioning is after the q th row. The estimator $\widehat{\beta}_n^{**}$ can be interpreted as a non-Gaussian counterpart of the 2SLS estimator of β , the latter being given by

$$\bar{\beta}_n = \frac{\bar{\Delta}_{1,n}' Q_{zz,n} \bar{\Delta}_{2,n}}{\bar{\Delta}_{2,n}' Q_{zz,n} \bar{\Delta}_{2,n}},$$

where $\bar{\Delta}_n = \left(\bar{\Delta}_{1,n}', \bar{\Delta}_{2,n}'\right)'$ and partitioning is after the q th row. The estimators $\widehat{\beta}_n^{**}$ 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 $\widehat{\beta}_n^{**}$ being based on an asymptotically efficient estimator (namely $\widehat{\Delta}_n^{**}/\sqrt{n}$) and $\bar{\beta}_n$ being based on the OLS estimator $\bar{\Delta}_n/\sqrt{n}$.

If Assumption 4SF holds, then

$$\sqrt{n}(\bar{\beta}_n - \beta) \rightarrow_d \mathcal{N}(0, \bar{\Sigma}_\beta), \quad \bar{\Sigma}_\beta = \left[\begin{pmatrix} 1 \\ -\beta \end{pmatrix}' \Omega \begin{pmatrix} 1 \\ -\beta \end{pmatrix} \right] (\pi' Q_{zz} \pi)^{-1}.$$

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*

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

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 estimator and Fuller's (1977) modification thereof). With non-Gaussian errors, on the other hand, the estimator $\hat{\beta}_n^{**}$ compares favorably with 2SLS whenever the inequality $\mathcal{I}^{-1} \leq \Omega$ 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

$$\hat{\beta}_n(\lambda_1, \lambda_2) = \frac{\hat{\Pi}_n(\lambda_1)' Q_{zz,n} \hat{\pi}_n(\lambda_2)}{\hat{\pi}_n(\lambda_2)' Q_{zz,n} \hat{\pi}_n(\lambda_2)}, \quad (\lambda_1, \lambda_2)' \in \mathbb{R}^2,$$

where

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

$$\left(\hat{\gamma}_1^{LAD}, \hat{\Pi}_n^{LAD} \right) = \arg \min_{(\gamma_1, \Pi)} \sum_{i=1}^n |y_{1i} - \gamma_1' x_i - \Pi' z_i|,$$

$$\left(\hat{\gamma}_2^{LAD}, \hat{\pi}_n^{LAD} \right) = \arg \min_{(\gamma_2, \pi)} \sum_{i=1}^n |y_{2i} - \gamma_2' x_i - \pi' z_i|.$$

In this notation $\hat{\beta}_n(0, 0)$ 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 $\hat{\beta}_n(\lambda_1, \lambda_2)$ is readily obtained (by means of the delta method) 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 $\hat{\beta}_n(\lambda_1, 0)$ is asymptotically equivalent to the 2SLAD(λ_1) estimator and that $\hat{\beta}_n(1, 1)$ is asymptotically equivalent to the D2SLAD estimator(s).

Because $(\hat{\Delta}_{1,n}^{**}/\sqrt{n}, \hat{\Delta}_{2,n}^{**}/\sqrt{n})$ is an asymptotically efficient estimator of (δ_1, δ_2) in (3), it compares favorably with $(\hat{\Pi}_n(\lambda_1), \hat{\pi}_n(\lambda_2))$ 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 $\hat{\beta}_n(\lambda_1, \lambda_2)$ (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.²⁰

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

$$\widehat{W}_n^{**} = \frac{(\hat{\beta}_n^{**})^2}{\hat{\Sigma}_\beta^{**}/n}, \quad \hat{\Sigma}_\beta^{**} = \left[\begin{pmatrix} 1 \\ -\hat{\beta}_n^{**} \end{pmatrix}' \hat{\mathcal{I}}_n^{*-1} \begin{pmatrix} 1 \\ -\hat{\beta}_n^{**} \end{pmatrix} \right] (\hat{\pi}_n' Q_{zz,n} \hat{\pi}_n)^{-1}. \quad (25)$$

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*

$$\widehat{W}_n^{**} = \frac{(S_n' Q_{zz}^{1/2} \pi)^2}{\pi' Q_{zz} \pi} + o_p(1) \rightarrow_d \chi^2(1; b^2 \mathcal{I}_{11.2} \pi' Q_{zz} \pi).$$

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

²⁰This result, and the other optimality results obtained under the assumption of strong identification, should generalize straightforwardly to models with multiple endogenous regressors. The distribution theory developed under weak identification can also be generalized to models with multiple endogenous regressors, but to the best of our knowledge the “near” optimality properties of Moreira (2003)-type inference procedures has not been established for models of this type.

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

It should be emphasized, however, that the displayed confidence interval is invalid (i.e., 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^{**} .²¹

Remarks. (i) The relation between $\hat{\beta}_n^{**}$ and the 2SLS estimator $\bar{\beta}_n$ can be further elucidated by noticing that

$$\hat{\beta}_n^{**} = \frac{\sum_{i=1}^n \hat{y}_{2i}^{**} \hat{y}_{1i}^{**}}{\sum_{i=1}^n \hat{y}_{2i}^{**} \hat{y}_{2i}^{**}}, \quad \hat{y}_{ji}^{**} = z_i' \hat{\Delta}_{j,n}^{**} / \sqrt{n},$$

a representation perfectly analogous to

$$\bar{\beta}_n = \frac{\sum_{i=1}^n \bar{y}_{2i} \bar{y}_{1i}}{\sum_{i=1}^n \bar{y}_{2i} \bar{y}_{2i}}, \quad \bar{y}_{ji} = z_i' \bar{\Delta}_{j,n} / \sqrt{n},$$

the latter being Basmann’s (1959) interpretation of 2SLS.

(ii) For the purposes of understanding the asymptotic efficiency of $\hat{\beta}_n^{**}$, it may be useful to recognize that it admits a minimum distance interpretation. Indeed,

$$\left(\hat{\Delta}_{2,n}^{**} / \sqrt{n}, \hat{\beta}_n^{**} \right) = \arg \min_{(\pi, \beta)} \begin{pmatrix} \hat{\Delta}_{1,n}^{**} / \sqrt{n} - \pi \beta \\ \hat{\Delta}_{2,n}^{**} / \sqrt{n} - \pi \end{pmatrix}' (\mathcal{I} \otimes Q_{zz,n}) \begin{pmatrix} \hat{\Delta}_{1,n}^{**} / \sqrt{n} - \pi \beta \\ \hat{\Delta}_{2,n}^{**} / \sqrt{n} - \pi \end{pmatrix},$$

a characterization which closely resembles the following (well known) minimum distance interpretation of $\bar{\beta}_n$:

$$\left(\hat{\pi}_n^{OLS}, \bar{\beta}_n \right) = \arg \min_{(\pi, \beta)} \begin{pmatrix} \bar{\Delta}_{1,n} / \sqrt{n} - \pi \beta \\ \bar{\Delta}_{2,n} / \sqrt{n} - \pi \end{pmatrix}' (\Omega^{-1} \otimes Q_{zz,n}) \begin{pmatrix} \bar{\Delta}_{1,n} / \sqrt{n} - \pi \beta \\ \bar{\Delta}_{2,n} / \sqrt{n} - \pi \end{pmatrix}.$$

²¹A “conditional” version of the Wald test will be valid also under weak identification, but will be no easier to implement than the test based on \widehat{LR}_n^{**} and will not be “nearly” efficient when identification is weak (Andrews, Moreira, and Stock (2006b)).

(iii) Recently, Hansen, McDonald, and Newey (2006) have proposed a nonlinear IV estimator of β based on a parametric family of densities for u_i in a model of the form

$$y_{1i} = \Gamma_1' x_i + \beta y_{2i} + u_i \quad (i = 1, \dots, n). \quad (26)$$

(That is, their model consists of the first equation of (1), but does contain a first stage equation relating the endogenous regressor to the instruments.) If the true density of u_i is a member of the family used for estimation purposes, Hansen, McDonald, and Newey's (2006) estimator is locally efficient at the model (1) under the assumptions of that paper. On the other hand, their estimator is inferior to $\hat{\beta}_n^{**}$ under the (stronger) assumptions made herein. (Their estimator is asymptotically equivalent to $\hat{\beta}_n^{**}$ when the errors u_i and v_{2i} are in (1) are mutually independent and/or Gaussian, but not in general.)

5. 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

$$\begin{aligned} \mathcal{L}_n(d, g) &= \sum_{i=1}^n \log f [y_{1i} - \gamma_{1n}(g_1)' x_i - \delta_{1n}(d_1)' z_i, y_{2i} - \gamma_{2n}(g_2)' x_i - \delta_{2n}(d_2)' z_i] \\ &\quad - \sum_{i=1}^n \log f [y_{1i} - \gamma_1' x_i - \delta_1' z_i, y_{2i} - \gamma_2' x_i - \delta_2' z_i] \end{aligned} \quad (27)$$

denote the log likelihood ratio function associated with the local reparameterization

$$\gamma = \begin{bmatrix} \gamma_{1n}(g_1) \\ \gamma_{2n}(g_2) \end{bmatrix} = \begin{bmatrix} \gamma_1 + g_1/\sqrt{n} \\ \gamma_2 + g_2/\sqrt{n} \end{bmatrix}, \quad \delta = \begin{bmatrix} \delta_{1n}(d_1) \\ \delta_{2n}(d_2) \end{bmatrix} = \begin{bmatrix} \delta_1 + d_1/\sqrt{n} \\ \delta_2 + d_2/\sqrt{n} \end{bmatrix}, \quad (28)$$

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 distributions associated with $(d, g) = (0, 0)$ ”, respectively, and let

$$\ell_i = \ell(y_{1i} - \gamma_1' x_i - \delta_1' z_i, y_{2i} - \gamma_2' x_i - \delta_2' z_i).$$

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

(a) *If Assumptions 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

$$\mathcal{L}_n^\delta(d_n) = d_n' (\mathcal{I} \otimes Q_{zz}) \Delta_n^\delta - \frac{1}{2} d_n' (\mathcal{I} \otimes Q_{zz}) d_n, \quad \Delta_n^\delta = (\mathcal{I}^{-1} \otimes Q_{zz}^{-1}) \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell_i \otimes z_i,$$

$$\Delta_n^\delta \rightarrow_{d_{\delta, \gamma}} \mathcal{N}(0, \mathcal{I}^{-1} \otimes Q_{zz}^{-1}).$$

(b) If, moreover, Assumptions 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

$$\begin{aligned} \mathcal{L}_n^\gamma(g_n) &= g_n' (\mathcal{I} \otimes Q_{xx}) \Delta_n^\gamma - \frac{1}{2} g_n' (\mathcal{I} \otimes Q_{xx}) g_n, \quad \Delta_n^\gamma = (\mathcal{I}^{-1} \otimes Q_{xx}^{-1}) \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell_i \otimes x_i, \\ &\left(\begin{array}{c} \Delta_n^\delta \\ \Delta_n^\gamma \end{array} \right) \rightarrow_{d_{\delta, \gamma}} \mathcal{N} \left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \mathcal{I}^{-1} \otimes Q_{zz}^{-1} & 0 \\ 0 & \mathcal{I}^{-1} \otimes Q_{xx}^{-1} \end{array} \right) \right]. \end{aligned}$$

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

$$\tilde{\delta}_n(\hat{\delta}_n, \hat{\gamma}_n) = \hat{\delta}_n + \frac{1}{\sqrt{n}} \hat{\Delta}_n^\delta(\hat{\delta}_n, \hat{\gamma}_n),$$

where

$$\hat{\Delta}_n^\delta(\hat{\delta}_n, \hat{\gamma}_n) = \left[\tilde{\mathcal{I}}_n(\hat{\delta}_n, \hat{\gamma}_n)^{-1} \otimes Q_{zz, n}^{-1} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_n(\hat{v}_i) \otimes z_i, \quad (29)$$

$$\tilde{\mathcal{I}}_n(\hat{\delta}_n, \hat{\gamma}_n) = n^{-1} \sum_{i=1}^n \hat{\ell}_n(\hat{v}_i) \hat{\ell}_n(\hat{v}_i)', \quad \hat{v}_i = \begin{pmatrix} y_{1i} - \hat{\gamma}'_{1n} x_i - \hat{\delta}'_{1n} z_i \\ y_{2i} - \hat{\gamma}'_{2n} x_i - \hat{\delta}'_{2n} z_i \end{pmatrix}, \quad (30)$$

and

$$\hat{\ell}_n(v) = -\frac{\partial \hat{f}_n(v) / \partial v}{\hat{f}_n(v) + a_n}, \quad \hat{f}_n(v) = \frac{1}{nh_n^2} \sum_{i=1}^n K\left(\frac{v - \hat{v}_i}{h_n}\right). \quad (31)$$

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

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

and

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

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

Proof of Theorems 2, 4, and 6. Theorems 2 and 4 (a) are special cases of Theorem 6 (a) and Theorem 4 (b) is a special case of Theorem 6 (b), so it suffices to prove Theorem 6.

Theorem 6 can be derived with the help of Theorem A.2 because

$$\hat{\Delta}_n^{**} = \sqrt{n} \tilde{\delta}_n \left(\hat{\delta}_n, \hat{\gamma}_n \right), \quad \hat{\mathcal{I}}_n^{**} = \tilde{\mathcal{I}}_n \left(\hat{\delta}_n, \hat{\gamma}_n \right),$$

where $\hat{\delta}_n = \left(\hat{\Pi}'_n, \hat{\pi}'_n \right)'$ and $\hat{\gamma}_n$ is as in the main text.

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

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

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

Proof of Theorem A.1. Define

$$R(v, \theta) = 2 \left[\sqrt{\frac{f(v - \theta)}{f(v)}} - 1 - \frac{1}{2} \theta' \ell(v) \right] 1[f(v) > 0], \quad v, \theta \in \mathbb{R}^2,$$

and

$$\bar{R}(\theta) = \frac{1}{4} \theta' \mathcal{I} \theta + \int_{\mathbb{R}^2} R(v, \theta) f(v) dv, \quad \theta \in \mathbb{R}^2.$$

If Assumption 2 holds, then

$$\sqrt{f(v - \theta)} - \sqrt{f(v)} = \frac{1}{2} \theta' \int_0^1 \ell(v - \theta t) \sqrt{f(v - \theta t)} dt, \quad \forall v, \theta \in \mathbb{R}^2$$

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

$$\lim_{\eta \downarrow 0} V(\eta) = 0, \quad V(\eta) = \sup_{\|\theta\| \leq \eta, \theta \neq 0} \|\theta\|^{-2} \int_{\mathbb{R}^2} R(v, \theta)^2 f(v) dv. \quad (32)$$

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

$$\lim_{\eta \downarrow 0} \bar{V}(\eta) = 0, \quad \bar{V}(\eta) = \sup_{\|\theta\| \leq \eta, \theta \neq 0} \|\theta\|^{-2} \bar{R}(\theta). \quad (33)$$

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

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

where

$$R_{i,n} = R \left[\begin{pmatrix} y_{1i} - \gamma_1' x_i - \delta_1' z_i \\ y_{2i} - \gamma_2' x_i - \delta_2' z_i \end{pmatrix}, \begin{pmatrix} d_{1n}' z_i / \sqrt{n} \\ d_{2n}' z_i / \sqrt{n} \end{pmatrix} \right], \quad \xi_{i,n} = \xi \left[d_n' \frac{\ell_i \otimes z_i}{\sqrt{n}} + R_{i,n} \right],$$

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:

$$\sum_{i=1}^n R_{i,n} = -\frac{1}{4} d_n' (\mathcal{I} \otimes Q_{zz}) d_n + o_{p_{\delta, \gamma}}(1), \quad (34)$$

$$\max_{1 \leq i \leq n} |\xi_{i,n}| = o_{p_{\delta, \gamma}}(1), \quad (35)$$

$$\sum_{i=1}^n \left[d_n' \frac{\ell_i \otimes z_i}{\sqrt{n}} + R_{i,n} \right]^2 = d_n' (\mathcal{I} \otimes Q_{zz}) d_n + o_{p_{\delta, \gamma}}(1). \quad (36)$$

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

Proof of (34). The random variables $R_{1,n}, \dots, R_{n,n}$ are independent and satisfy

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

where the penultimate equality uses (32) and Assumption 1(a). As a consequence,

$$\sum_{i=1}^n R_{i,n} = \sum_{i=1}^n \mathbb{E}(R_{i,n}) + o_p(1),$$

where

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

By Assumption 1(a),

$$d'_n (\mathcal{I} \otimes Q_{zz,n}) d_n = d'_n (\mathcal{I} \otimes Q_{zz}) d_n + o(1).$$

Moreover,

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

where where the penultimate equality uses (33) and Assumption 1(a).

Proof of (35). Because $\lim_{t \rightarrow 0} \xi(t) = 0$ (by Taylor's Theorem), the result follows from the fact that

$$\max_{1 \leq i \leq n} \left\| \frac{\ell_i \otimes z_i}{\sqrt{n}} \right\| = o_p(1)$$

and

$$\max_{1 \leq i \leq n} |R_{i,n}| \leq \sqrt{\sum_{i=1}^n R_{i,n}^2} = o_p(1),$$

where the first convergence result uses $\ell_i \sim i.i.d. (0, \mathcal{I})$ and Assumption 1(a), while the second convergence result uses the relation $\mathbb{E}(\sum_{i=1}^n R_{i,n}^2) = o(1)$ established in the proof of (34).

Proof of (36). Because $\sum_{i=1}^n R_{i,n}^2 = o_p(1)$ and

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

it suffices to show that

$$\frac{1}{n} \sum_{i=1}^n \ell_i \ell'_i \otimes z_i z'_i = \mathcal{I} \otimes Q_{zz} + o_p(1).$$

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 $(\hat{\delta}_n, \hat{\gamma}_n)$ that we may assume $(\hat{\delta}_n, \hat{\gamma}_n) = (\delta, \gamma)$. (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

$$\check{\Delta}_n^\delta = [\check{\mathcal{I}}_n^{-1} \otimes Q_{zz,n}^{-1}] \frac{1}{\sqrt{n}} \sum_{i=1}^n \check{\ell}_n(v_i) \otimes z_i = \Delta_n^\delta + o_p(1) \quad (37)$$

and

$$\check{\mathcal{I}}_n = n^{-1} \sum_{i=1}^n \check{\ell}_n(v_i) \check{\ell}_n(v_i)' = \mathcal{I} + o_p(1), \quad (38)$$

where

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

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

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

It follows from (the proof of) Lemma 3.1 and Remark 3.2 of Schick (1987) that condition (37) is implied by condition (38), Assumptions 1(a) and 2, and the following conditions:²²

$$\mathbb{E} \left[\int_{\mathbb{R}^2} \|\check{\ell}_n(v) - \ell(v)\|^2 f(v) dv \right] = o(1), \quad (39)$$

$$\max_{1 \leq i \leq n} \mathbb{E} \left[\int_{\mathbb{R}^2} \|\check{\ell}_n(v) - \check{\ell}_{n,i}(v)\|^2 f(v) dv \right] = o\left(\frac{1}{n}\right). \quad (40)$$

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

$$\int_{\mathbb{R}^2} \left\| -\frac{\partial f_n(v)/\partial v}{f_n(v) + a_n} - \ell(v) \right\|^2 f(v) dv = o(1), \quad (41)$$

where $f_n(v) = \int_{\mathbb{R}^2} f(v - h_n r) K(r) dr = \mathbb{E}[\check{f}_n(v)]$. It follows from this result that

$$\int_{\mathbb{R}^2} \left\| \frac{\partial f_n(v)/\partial v}{f_n(v) + a_n} \right\|^2 f(v) dv = O(1). \quad (42)$$

Now, using Assumptions 2 and 5, we have

$$\sup_{v \in \mathbb{R}^2} \mathbb{E} \left[\|\check{f}_n(v) - f_n(v)\|^2 \right] = O\left(\frac{1}{nh_n^2}\right)$$

and

²²Conditions (39) and (40) are counterparts of Schick's (1987) conditions (3.2) and (3.6). No counterpart of Schick's (1987) condition (3.1) is needed because $\sum_{i=1}^n z_i = 0$. Also, the present definition of $\check{\ell}_{n,i}$ ensures that $\check{\ell}_{n,i}(v_i) = \check{\ell}_n(v_i)$ for every i , implying in particular that the natural counterpart of Schick's (1987) conditions (3.5) is satisfied.

$$\sup_{v \in \mathbb{R}^2} \mathbb{E} \left[\left\| \frac{\partial \check{f}_n(v)}{\partial v} - \frac{\partial f_n(v)}{\partial v} \right\|^2 \right] = O \left(\frac{1}{nh_n^4} \right).$$

Utilizing these facts, (42), and the decomposition

$$\frac{\frac{\partial \check{f}_n(v)}{\partial v}}{\check{f}_n(v) + a_n} - \frac{\frac{\partial f_n(v)}{\partial v}}{f_n(v) + a_n} = -\frac{\frac{\partial f_n(v)}{\partial v} \check{f}_n(v) - f_n(v)}{f_n(v) + a_n} + \frac{\frac{\partial \check{f}_n(v)}{\partial v} - \frac{\partial f_n(v)}{\partial v}}{\check{f}_n(v) + a_n},$$

it is easily shown that

$$\int_{\mathbb{R}^2} \mathbb{E} \left[\left\| \frac{\frac{\partial \check{f}_n(v)}{\partial v}}{\check{f}_n(v) + a_n} - \frac{\frac{\partial f_n(v)}{\partial v}}{f_n(v) + a_n} \right\|^2 \right] f(v) dv = O \left(\frac{1}{na_n^2 h_n^4} \right) = o(1), \quad (43)$$

a result which can be combined with (41) to yield (39).

It follows from (42) – (43) that

$$\int_{\mathbb{R}^2} \mathbb{E} \left[\left\| \frac{\frac{\partial \check{f}_n(v)}{\partial v}}{\check{f}_n(v) + a_n} \right\|^2 \right] f(v) dv = O(1).$$

Utilizing this fact, Assumption 5, and the decomposition

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

it is easily shown that (40) holds.

Finally, condition (38) holds because

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

where the first equality uses the fact that $\check{\ell}_{n,i}(v_i) = \check{\ell}_n(v_i)$ for each i and the second equality uses (41) and (43). ■

Remark. With the possible exception of (41), all steps in the proof of Theorem A.2 remain valid if the condition $\sup_{r \in \mathbb{R}} |k'(r)|/k(r) < \infty$ of Assumption 5(a) is replaced by the condition $\int_{\mathbb{R}} k'(r)^2 dr < \infty$. 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 $\sup_{v \in \mathbb{R}^2} \|\dot{f}(v)\| < \infty$, then (41) is satisfied (for any kernel) provided $\overline{\lim}_{n \rightarrow \infty} h_n/a_n < \infty$. This is so because

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

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

$$\int_{\mathbb{S}_f} \left[\sqrt{f(v)} - \sqrt{f_n(v)} \right]^2 dv = o(h_n^2), \quad (44)$$

$$\int_{\mathbb{S}_f} \left\| \frac{\partial f_n(v)/\partial v}{f_n(v) + a_n} \sqrt{f_n(v)} - \frac{\dot{f}(v)}{f(v)} \sqrt{f(v)} \right\|^2 dv = o(1), \quad (45)$$

and the bound

$$\sup_{v \in \mathbb{R}^2} \left\| \frac{\partial f_n(v)/\partial v}{f_n(v) + a_n} \right\|^2 \leq \left(\sup_{v \in \mathbb{R}^2} \|\dot{f}(v)\| \right)^2 / a_n^2.$$

(The result (44) can be shown by means of Proposition A.7 of Koul and Schick (1996), while (45) 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., AND H. RUBIN (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., AND V. MARMER (2005): “Exactly Distribution-Free Inference in Instrumental Variables Regression with Possibly Weak Instruments,” Working Paper, Yale University.
- ANDREWS, D. W. K., M. J. MOREIRA, AND J. H. STOCK (2006a): “Optimal Two-sided Invariant Similar Tests for Instrumental Variables Regression,” *Econometrica*, 74, 715–752.
- (2006b): “Performance of Conditional Wald Tests in IV Regression with Weak Instruments,” *Journal of Econometrics*, forthcoming.
- ANDREWS, D. W. K., AND G. SOARES (2006): “Rank Tests for Instrumental Variables Regression with Weak Instruments,” Working Paper, Yale University.
- ANDREWS, D. W. K., AND J. H. STOCK (2006a): “Inference with Weak Instruments,” *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, forthcoming.
- (2006b): “Testing with Many Weak Instruments,” *Journal of Econometrics*, forthcoming.
- BASMANN, R. L. (1959): “The Computation of Generalized Classical Estimates of Coefficients in a Structural Equation,” *Econometrica*, 27, 72–81.
- BEKKER, P. A. (1994): “Alternative Approximations to the Distributions of Instrumental Variables Estimators,” *Econometrica*, 62, 657–681.
- BICKEL, P. J. (1982): “On Adaptive Estimation,” *Annals of Statistics*, 10, 647–671.
- BOUND, J., D. A. JAEGER, AND R. M. BAKER (1995): “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak,” *Journal of the American Statistical Association*, 90, 443–450.
- CHAO, J. C., AND N. R. SWANSON (2005): “Consistent Estimation With a Large Number of Weak Instruments,” *Econometrica*, 73, 1673–1692.

- CHIODA, L., AND M. JANSSON (2006): “Optimal Invariant Inference When the Number of Instruments is Large,” Working Paper, Princeton University.
- CHOI, S., W. J. HALL, AND A. SCHICK (1996): “Asymptotically Uniformly Most Powerful Tests in Parametric and Semiparametric Models,” *Annals of Statistics*, 24, 841–861.
- COX, D. R., AND N. R. REID (1987): “Parameter Orthogonality and Approximate Conditional Inference (with Discussion),” *Journal of the Royal Statistical Society, Series B*, 49, 1–39.
- DUDLEY, R. M. (2002): *Real Analysis and Probability*. New York: Cambridge University Press.
- DUFOUR, J.-M. (1997): “Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models,” *Econometrica*, 65, 1365–1387.
- (2003): “Identification, Weak Instruments, and Statistical Inference in Econometrics,” *Canadian Journal of Economics*, 36, 767–808.
- FARAWAY, J. J. (1992): “Smoothing in Adaptive Estimation,” *Annals of Statistics*, 20, 414–427.
- FULLER, W. A. (1977): “Some Properties of a Modification of the Limited Information Estimator,” *Econometrica*, 45, 939–953.
- HAHN, J. (2002): “Optimal Inference with Many Instruments,” *Econometric Theory*, 18, 140–168.
- HAHN, J., AND J. HAUSMAN (2003): “Weak Instruments: Diagnosis and Cures in Empirical Econometrics,” *American Economic Review*, 93, 118–125.
- HANSEN, C., J. HAUSMAN, AND W. NEWEY (2005): “Estimation with Many Instrumental Variables,” Working Paper, University of Chicago GSB.
- HANSEN, C., J. B. McDONALD, AND W. K. NEWEY (2006): “Instrumental Variables Estimation with Flexible Distributions,” Working Paper, University of Chicago GSB.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 1029–1054, 50.
- JANSSON, M. (2006): “Semiparametric Power Envelopes for Tests of the Unit Root Hypothesis,” Working Paper, UC Berkeley.

- JIN, K. (1992): “Empirical Smoothing Parameter Selection in Adaptive Estimation,” *Annals of Statistics*, 20, 1844–1874.
- KLEIBERGEN, F. (2002): “Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression,” *Econometrica*, 70, 1781–1803.
- KOENKER, R., AND G. BASSETT (1978): “Regression Quantiles,” *Econometrica*, 46, 33–50.
- KOUL, H. L., AND A. SCHICK (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.
- NELSON, C. R., AND R. STARTZ (1990a): “The Distribution of the Instrumental Variables Estimator and its t -Ratio When the Instrument is a Poor One,” *Journal of Business*, 63, S125–S140.
- (1990b): “Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator,” *Econometrica*, 58, 967–976.
- NEWAY, W. K. (1988): “Adaptive Estimation of Regression Models Via Moment Restrictions,” *Journal of Econometrics*, 38, 301–339.
- (2004): “Efficient Semiparametric Estimation Via Moment Restrictions,” *Econometrica*, 72, 1877–1897.
- POLLARD, D. (1997): “Another Look at Differentiability in Quadratic Mean,” in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, ed. by D. Pollard, E. Torgersen, and G. L. Yang. New York: Springer-Verlag, 305–314.
- PORTNOY, S., AND R. KOENKER (1989): “Adaptive L -Estimation for Linear Models,” *Annals of Statistics*, 17, 362–381.
- POWELL, J. L. (1983): “The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators,” *Econometrica*, 51, 1569–1575.
- PUHALSKII, A., AND V. SPOKOINY (1998): “On Large-Deviation Efficiency in Statistical Inference,” *Bernoulli*, 4, 203–272.
- SCHICK, A. (1987): “A Note on the Construction of Asymptotically Linear Estimators,” *Journal of Statistical Planning and Inference*, 16, 89–105.

———— (1993): “On Efficient Estimation in Regression Models,” *Annals of Statistics*, 21, 1486–1521.

———— (1997): “Efficient Estimates in Linear and Nonlinear Regression with Heteroscedastic Errors,” *Journal of Statistical Planning and Inference*, 58, 371–387.

STAIGER, D., AND J. H. STOCK (1997): “Instrumental Variables Estimation with Weak Instruments,” *Econometrica*, 65, 557–586.

STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments,” *Journal of Business and Economic Statistics*, 20, 518–529.

STOCK, J. H., AND M. YOGO (2005): “Asymptotic Distributions of Instrumental Variables Statistics with Many Weak Instruments,” in *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. by D. W. K. Andrews, and J. H. Stock. New York: Cambridge University Press, 109-120.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. New York: Cambridge University Press.