Econometrica Supplementary Material

SUPPLEMENT TO "ENTROPIC LATENT VARIABLE INTEGRATION VIA SIMULATION"
(*Econometrica*, Vol. 82, No. 1, January 2014, 345–385)

BY SUSANNE M. SCHENNACH

INTRODUCTION

THIS SUPPLEMENTAL MATERIAL INCLUDES (i) proofs omitted from the main text, (ii) additional simulation examples, (iii) an extended notion of the identified set, (iv) difficulties associated with the use of alternative discrepancies, (v) inference methods, (vi) computational details of the implementation of the method in the paper, (vii) an example of equivalence to standard bounding techniques, and (viii) relationships with earlier information-theoretic and entropy-based methods.

APPENDIX B: PROOFS OMITTED FROM THE MAIN TEXT

Throughout the proofs, we denote $\rho(u|z;\theta)$ by $\rho(u|z)$, making the dependence on θ implicit (as all arguments hold pointwise in θ).

PROOF OF PROPOSITION 2.1: This proof frequently makes the use of random variables (and expectations, probabilities, or support thereof) that are conditional on the event $Z = z$ and the following qualifications will apply throughout. Since we consider regular conditional probability measures, a distribution (say, of a random variable U) conditional on $Z = z$ will be a well defined probability measure for all z in a set $\mathcal{Z}' \subseteq \mathcal{Z}$ of probability 1 under the distribution of Z. All statements for a given z will be for $z \in \mathcal{Z}'$ and we need not consider $z \in \mathcal{Z} \setminus \mathcal{Z}'$, since such events have probability 0 and will not affect any unconditional probabilities or expectations. Also, recall that $g(u, z, \theta)$ is assumed to be measurable throughout. Finally, since all arguments hold pointwise in θ, we make dependence on θ implicit and denote $\rho(u|z;\theta)$ by $\rho(u|z)$, $\lambda(u|z;\theta)$ by $\lambda(u|z)$, and $\dot{u}(z,\theta)$ by $\dot{u}(z)$.

We now verify that the example satisfies the conditions of Definition 2.2. We first note that the support of $\lambda(\cdot|z)$ is $\mathcal{U}$ by construction and that ρ differs from λ only by a multiplicative prefactor $C(z,\theta)\exp(-\|g(u,z,\theta) - g(\dot{u}(z), z, \theta)\|^2)$. Since $C(z,\theta) \geq 1$ by construction, the prefactor is nonvanishing for any finite $g(u, z, \theta)$ and it follows that the supports of $\rho(\cdot|z)$ and $\lambda(\cdot|z)$ agree.

Next, we check the differentiability requirement on $E_\pi[\ln E_\rho[\exp(\gamma' g(U, Z, \theta))|Z]]$. We first check that the second derivative is finite—the boundedness of the function itself and of its first derivative follow by similar arguments. By the same reasoning as in the proof of Lemma A.1, the interchanges of derivatives

DOI: 10.3982/ECTA9748

with expectation performed below are allowed. We then have

$$\begin{aligned}
&\frac{\partial^2}{\partial\gamma\,\partial\gamma'}E_\pi\big[\ln E_\rho\big[\exp\big(\gamma' g(U,Z,\theta)\big)|Z\big]\big] \\
&\quad = E_\pi\left[\frac{E_\rho[g(U,Z,\theta)g'(U,Z,\theta)\exp(\gamma' g(U,Z,\theta))|Z]}{E_\rho[\exp(\gamma' g(U,Z,\theta))|Z]}\right] \\
&\qquad - E_\pi\big[\tilde{g}(Z,\theta,\gamma)\tilde{g}'(Z,\theta,\gamma)\big] \\
&\quad = E_\pi\Big[E_\rho\big[\big(g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\big)\big(g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\big)' \\
&\qquad \times \exp\big(\gamma' g(U,Z,\theta)\big)|Z\big] \\
&\qquad /E_\rho\big[\exp\big(\gamma' g(U,Z,\theta)\big)|Z\big]\Big],
\end{aligned} \tag{S.1}$$

where

$$\tilde{g}(z,\theta,\gamma) \equiv \frac{E_\rho[g(U,Z,\theta)\exp(\gamma' g(U,Z,\theta))|Z=z]}{E_\rho[\exp(\gamma' g(U,Z,\theta))|Z=z]}.$$

We then bound each element of the matrix (S.1) by a single scalar quantity: For $i,j\in\{1,\dots,d_g\}$, we have

$$\begin{aligned}
&E_\rho\big[\big(g_i(U,Z,\theta)-\tilde{g}_i(Z,\theta,\gamma)\big)\big(g_j(U,Z,\theta)-\tilde{g}_j(Z,\theta,\gamma)\big) \\
&\qquad \times \exp\big(\gamma' g(U,Z,\theta)\big)|Z=z\big] \\
&\qquad /E_\rho\big[\exp\big(\gamma' g(U,Z,\theta)\big)|Z=z\big] \\
&\quad \le \Big(E_\rho\big[\big(g_i(U,Z,\theta)-\tilde{g}_i(Z,\theta,\gamma)\big)^2\exp\big(\gamma' g(U,Z,\theta)\big)|Z=z\big] \\
&\qquad /E_\rho\big[\exp\big(\gamma' g(U,Z,\theta)\big)|Z=z\big]\Big)^{1/2} \\
&\qquad \times \Big(E_\rho\big[\big(g_j(U,Z,\theta)-\tilde{g}_j(Z,\theta,\gamma)\big)^2\exp\big(\gamma' g(U,Z,\theta)\big)|Z=z\big] \\
&\qquad /E_\rho\big[\exp\big(\gamma' g(U,Z,\theta)\big)|Z=z\big]\Big)^{1/2} \\
&\quad \le \frac{E_\rho[\|g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\|^2\exp(\gamma' g(U,Z,\theta))|Z=z]}{E_\rho[\exp(\gamma' g(U,Z,\theta))|Z=z]} \\
&\quad = \frac{E_\rho[\|g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\|^2\exp(\gamma' g(U,Z,\theta))|Z=z]}{E_\rho[\exp(\gamma' g(U,Z,\theta))|Z=z]} \\
&\qquad \times \frac{\exp(-\gamma' g(\dot{u}(z),z,\theta))}{\exp(-\gamma' g(\dot{u}(z),z,\theta))}
\end{aligned} \tag{S.2}$$

$$
\begin{aligned}
&= E_\rho\big[\big\|g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\big\|^2 \\
&\quad \times \exp\big(\gamma'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big] \\
&\quad /E_\rho\big[\exp\big(\gamma'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big],
\end{aligned}
$$

where we have (i) used the Cauchy–Schwarz inequality, (ii) used the fact that $(g_i(u,z,\theta)-\tilde{g}_i(z,\theta,\gamma))^2 \le \|g(u,z,\theta)-\tilde{g}(z,\theta,\gamma)\|^2$ for $i=1,\dots,d_g$, and (iii) multiplied the numerator and denominator by the same nonvanishing factor $\exp(-\gamma' g(\dot{u}(z),z,\theta))$.

We now bound, in turn, the numerator and the denominator of (S.2). Since the expected square deviation about the mean is less than about any other point (such as $g(\dot{u}(z),z,\theta)$), we have

$$
\begin{aligned}
&E_\rho\big[\big\|g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\big\|^2 \\
&\quad \times \exp\big(\gamma'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big] \\
&\le E_\rho\big[\big\|g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big\|^2 \\
&\quad \times \exp\big(\gamma'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big].
\end{aligned}
$$

Next, since a polynomial can be bounded by a suitable linear combination of exponentials (uniformly for any value of their corresponding argument),

$$
\begin{aligned}
\text{(S.3)}\quad &E_\rho\big[\big\|g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big\|^2 \\
&\quad \times \exp\big(\gamma'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big] \\
&\le \sum_{j=0}^{d_g} A_j E_\rho\big[\exp\big(\gamma_j'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big]
\end{aligned}
$$

for some finite $A_0,\dots,A_{d_g} \in \mathbb{R}^+$ and some $\gamma_0,\dots,\gamma_{d_g}$, each taking value in $\mathbb{R}^{d_g}$ and lying in an ε-neighborhood of γ (for some finite $\varepsilon > 0$ independent of z). Considering any one term in the sum (S.3), we have, by the definition of ρ,

$$
\begin{aligned}
\text{(S.4)}\quad &A_j E_\rho\big[\exp\big(\gamma_j'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big)\big)|Z=z\big] \\
&= A_j\Big(E_\lambda\big[\exp\big(\gamma_j'\big(g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big) \\
&\quad -\big\|g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big\|^2\big)|Z=z\big] \\
&\quad /E_\lambda\big[\exp\big(-\big\|g(U,Z,\theta)-g\big(\dot{u}(z),z,\theta\big)\big\|^2\big)|Z=z\big]\Big),
\end{aligned}
$$

where the denominator is simply the reciprocal of the normalization constant $C(z,\theta)$. The denominator of (S.4) can be easily bounded below by exploiting the assumed presence, in λ, of a point mass of probability $q > 0$ at $U = \dot{u}(z)$,

$$
\begin{aligned}
(\text{S.5}) \qquad & E_\lambda\big[\exp\big(-\big\|g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big\|^2\big)|Z = z\big] \\
&\geq E_\lambda\big[\exp\big(-\big\|g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big\|^2\big) \\
&\quad \times 1\big(g(U, Z, \theta) = g\big(\dot{u}(z), z, \theta\big)\big)|Z = z\big] \\
&= E_\lambda\big[\exp(0)1\big(g(U, Z, \theta) = g\big(\dot{u}(z), z, \theta\big)\big)|Z = z\big] \\
&= E_\lambda\big[1\big(g(U, Z, \theta) = g\big(\dot{u}(z), z, \theta\big)\big)|Z = z\big] \\
&\geq E_\lambda\big[1\big(U = \dot{u}(z)\big)|Z = z\big] = q > 0,
\end{aligned}
$$

where we have used the fact that (i) including an indicator function multiplier in an expectation of a positive quantity can only reduce its value (in a slight abuse of notation, we take the convention that the expectation of an indicator function turning on at the location of a point mass simply yields the probability mass assigned to that point) and (ii) the event $g(U, Z, \theta) = g(\dot{u}(z), z, \theta)$ is no less probable than $U = \dot{u}(z)$ because there may be multiple $u \in \mathcal{U}$ such that $g(u, z, \theta) = g(\dot{u}(z), z, \theta)$. We can then bound (S.4) as

$$
\begin{aligned}
(\text{S.6}) \qquad & A_j E_\rho\big[\exp\big(\gamma_j'\big(g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big)\big)|Z = z\big] \\
&\leq q^{-1} A_j E_\lambda\big[\exp\big(\gamma_j'\big(g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big) \\
&\quad - \big\|g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big\|^2\big)|Z = z\big] \\
&\leq q^{-1} A_j E_\lambda\big[\exp\big(\|\gamma_j\|\,\big\|g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big\| \\
&\quad - \big\|g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big\|^2\big)|Z = z\big] \\
&\leq q^{-1} A_j E_\lambda\Big[\exp\Big(\sup_{x\in\mathbb{R}}\big(\|\gamma_j\|x - x^2\big)\Big)\Big|Z = z\Big] \\
&= q^{-1} A_j E_\lambda\big[\exp\big(\|\gamma_j\|^2/4\big)|Z = z\big] \\
&= q^{-1} A_j \exp\big(\|\gamma_j\|^2/4\big) \\
&\leq q^{-1} A_j \exp\big(\big(\|\gamma\| + \varepsilon\big)^2/4\big),
\end{aligned}
$$

where we have used (i) inequality (S.5), (ii) the fact that $\sup_{x\in\mathbb{R}}(\|\gamma_j\|x - x^2) = \|\gamma_j\|^2/4$, and (iii) the fact that γ_j is in an ε-neighborhood of γ, combined with the triangle inequality.

We now obtain a lower bound on the denominator of (S.2),

$$
\begin{aligned}
(\text{S.7}) \qquad & E_\rho\big[\exp\big(\gamma'\big(g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big)\big)|Z = z\big] \\
&= E_\lambda\big[\exp\big(\gamma'\big(g(U, Z, \theta) - g\big(\dot{u}(z), z, \theta\big)\big)\big)
\end{aligned}
$$

$$
\begin{aligned}
&\quad \times \exp\bigl(-\bigl\|g(U,Z,\theta)-g\bigl(\dot{u}(z),z,\theta\bigr)\bigr\|^2\bigr)|Z=z\bigr] \\
&\quad /E_\lambda\bigl[\exp\bigl(-\bigl\|g(U,Z,\theta)-g\bigl(\dot{u}(z),z,\theta\bigr)\bigr\|^2\bigr)|Z=z\bigr] \\
&\geq E_\lambda\bigl[\exp\bigl(\gamma'\bigl(g(U,Z,\theta)-g\bigl(\dot{u}(z),z,\theta\bigr)\bigr)\bigr) \\
&\quad \times \exp\bigl(-\bigl\|g(U,Z,\theta)-g\bigl(\dot{u}(z),z,\theta\bigr)\bigr\|^2\bigr)|Z=z\bigr] \\
&\geq E_\lambda\bigl[\exp\bigl(\gamma'\bigl(g(U,Z,\theta)-g\bigl(\dot{u}(z),z,\theta\bigr)\bigr)\bigr) \\
&\quad \times \exp\bigl(-\bigl\|g(U,Z,\theta)-g\bigl(\dot{u}(z),z,\theta\bigr)\bigr\|^2\bigr) \\
&\quad \times 1\bigl(g(U,Z,\theta)=g\bigl(\dot{u}(z),z,\theta\bigr)\bigr)|Z=z\bigr] \\
&= E_\lambda\bigl[1\bigl(g(U,Z,\theta)=g\bigl(\dot{u}(z),z,\theta\bigr)\bigr)|Z=z\bigr] \\
&\geq E_\lambda\bigl[1(U=\dot{u})|Z=z\bigr]=q,
\end{aligned}
$$

where we have used the definition of ρ and the facts (i) that $E_\lambda[\exp(-\|g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\|^2)|Z=z] \leq E_\lambda[1|Z=z]=1$, (ii) that a multiplicative indicator function can only reduce the value of an expectation of a positive quantity, (iii) that $g(U,Z,\theta)=g(\dot{u}(z),z,\theta)$ implies that both exponentials equal 1, (iv) that $u=\dot{u}(z)$ implies $g(u,z,\theta)=g(\dot{u}(z),z,\theta)$, and (v) that the event $U=\dot{u}(z)$ given $Z=z$ has probability q by construction.

Combining the bounds (S.6) and (S.7), both of which hold uniformly in z and do not depend on z, the expectation in (S.1) can be bounded by a finite quantity at any $\gamma \in \mathbb{R}^{d_g}$,

$$
\begin{aligned}
\text{(S.8)}\qquad & E_\pi\left[\frac{E_\rho[\|g(U,Z,\theta)-\tilde{g}(Z,\theta,\gamma)\|^2\exp(\gamma' g(U,Z,\theta))|Z]}{E_\rho[\exp(\gamma' g(U,Z,\theta))|Z]}\right] \\
& \leq \bar{A}\exp\bigl(\bigl(\|\gamma\|+\varepsilon\bigr)^2/4\bigr),
\end{aligned}
$$

where $\bar{A} \equiv q^{-2}\sum_{j=0}^{d_g} A_j$. This bound on the second derivative of

$$E_\pi\bigl[\ln E_\rho\bigl[\exp\bigl(\gamma' g(U,Z,\theta)\bigr)|Z\bigr]\bigr] \equiv \tilde{M}(\gamma)$$

also implies that $\tilde{M}(\gamma)$ and $\partial\tilde{M}(\gamma)/\partial\gamma$ are finite at all $\gamma \in \mathbb{R}^{d_g}$. Indeed, $\partial\tilde{M}(\gamma)/\partial\gamma$ is given by the path integral

$$
\begin{aligned}
\text{(S.9)}\qquad \frac{\partial\tilde{M}(\gamma_1)}{\partial\gamma} &= \left.\frac{\partial\tilde{M}(\gamma_1)}{\partial\gamma}\right|_{\gamma_1=0} + \int_0^{\gamma_1}\frac{\partial^2\tilde{M}(\gamma)}{\partial\gamma\,\partial\gamma'}\cdot d\gamma \\
&= \left.\frac{\partial\tilde{M}(\gamma_1)}{\partial\gamma}\right|_{\gamma_1=0} + \int_0^{1}\frac{\partial^2\tilde{M}(\alpha\gamma_1)}{\partial\gamma\,\partial\gamma'}\cdot\gamma_1\,d\alpha,
\end{aligned}
$$

where we take a linear integration path for simplicity. Note that $\partial\tilde{M}(\gamma_1)/\partial\gamma|_{\gamma_1=0}$ is given by

$$\begin{aligned}
&E_\pi\left[\frac{E_\rho[g(U,Z,\theta)\exp(\gamma_1' g(U,Z,\theta))|Z]}{E_\rho[\exp(\gamma_1' g(U,Z,\theta))|Z]}\right]\Bigg|_{\gamma_1=0} \\
&\quad = E_\pi\big[E_\rho\big[g(U,Z,\theta)|Z\big]\big] \\
&\quad = E_\pi\left[\frac{E_\lambda[g(U,Z,\theta)\exp(-\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|^2)|Z]}{E_\lambda[\exp(-\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|^2)|Z]}\right] \\
&\quad = E_\pi\Big[E_\lambda\big[\big(g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big) \\
&\qquad\quad \times\exp\big(-\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\|^2\big)|Z\big] \\
&\qquad\quad /E_\lambda\big[\exp\big(-\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\|^2\big)|Z\big]\Big] \\
&\qquad\quad + E_\pi\big[E_\lambda\big[g\big(\dot{u}(Z),Z,\theta\big)\big]\big]
\end{aligned}$$

so that $\|\partial\tilde{M}(\gamma_1)/\partial\gamma|_{\gamma_1=0}\|$ is bounded by

$$\begin{aligned}
&E_\pi\Big[E_\lambda\big[\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\| \\
&\qquad \times\exp\big(-\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\|^2\big)|Z\big] \\
&\qquad /E_\lambda\big[\exp\big(-\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\|^2\big)|Z\big]\Big] \\
&\qquad + E_\pi\big[E_\lambda\big[\big\|g\big(\dot{u}(Z),Z,\theta\big)\big\|\big]\big] \\
&\quad \le E_\pi\big[q^{-1}E_\lambda\big[\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\| \\
&\qquad \times\exp\big(-\big\|g(U,Z,\theta)-g\big(\dot{u}(Z),Z,\theta\big)\big\|^2\big)|Z\big]\big] \\
&\qquad + E_\pi\big[\big\|g\big(\dot{u}(Z),Z,\theta\big)\big\|\big] \\
&\quad \le E_\pi\Big[q^{-1}E_\lambda\Big[\Big(\sup_{x\in\mathbb{R}}|x|\exp\big(-x^2\big)\Big)|Z\Big]\Big] + E_\pi\big[\big\|g\big(\dot{u}(Z),Z,\theta\big)\big\|\big] \\
&\quad \le E_\pi\big[q^{-1}E_\lambda[1|Z]\big] + E_\pi\big[\big\|g\big(\dot{u}(Z),Z,\theta\big)\big\|\big] \\
&\quad = q^{-1} + E_\pi\big[\big\|g\big(\dot{u}(Z),Z,\theta\big)\big\|\big] \\
&\quad \le q^{-1} + E_\pi\Big[\inf_{u\in\mathcal{U}}\big\|g(u,Z,\theta)\big\|\Big] + \omega \\
&\quad \le q^{-1} + E_{\mu\times\pi}\big[\big\|g(U,Z,\theta)\big\|\big] + \omega < \infty,
\end{aligned} \tag{S.10}$$

where we have used the facts that (i) result (S.5) holds, (ii) that $\sup_{x\in\mathbb{R}}|x|\times\exp(-x^2)\le 1$, (iii) that $\|g(\dot{u}(z),z,\theta)\|\le\inf_{u\in\mathcal{U}}\|g(u,z,\theta)\|+\omega$ by construction, and (iv) that $\inf_{u\in\mathcal{U}} \|g(u, Z, \theta)\| \leq \|g(\tilde{u}, Z, \theta)\|$ for any $\tilde{u} \in \mathcal{U}$, and that $E_{\mu\times\pi}[\|g(U, Z, \theta)\|]$ (with μ denoting the true data generating process of U given Z) must be finite for the model to be well defined. Combining (S.9), (S.10), and (S.8), we then have

$$\begin{aligned}
&\left\|\frac{\partial\tilde{M}(\gamma_1)}{\partial\gamma}\right\| \\
&\quad\leq \left\|\left.\frac{\partial\tilde{M}(\gamma_1)}{\partial\gamma}\right|_{\gamma_1=0}\right\| + \int_0^1 \left\|\frac{\partial^2\tilde{M}(\alpha\gamma_1)}{\partial\gamma\,\partial\gamma'}\right\| \|\gamma_1\|\,d\alpha \\
&\quad\leq q^{-1} + E_{\mu\times\pi}\big[\|g(U, Z, \theta)\|\big] + \omega + \|\gamma_1\| \sup_{\alpha\in[0,1]} \left\|\frac{\partial^2\tilde{M}(\alpha\gamma_1)}{\partial\gamma\,\partial\gamma'}\right\| \\
&\quad\leq q^{-1} + E_{\mu\times\pi}\big[\|g(U, Z, \theta)\|\big] + \omega + \|\gamma_1\|\bar{A}\exp\big((\|\gamma_1\| + \varepsilon)^2/4\big).
\end{aligned}$$

By a similar reasoning, $\tilde{M}(\gamma)$ is also bounded at each $\gamma \in \mathbb{R}^{d_g}$ since $\tilde{M}(\gamma_1) = \tilde{M}(0) + \int_0^1 \frac{\partial\tilde{M}(\alpha\gamma_1)}{\partial\gamma} \cdot \gamma_1\,d\alpha$ and $\tilde{M}(0) = 0$.

We have thus shown that the ρ provided satisfies the required support condition, and the corresponding $E_\pi[\ln E_\rho[\exp(\gamma' g(U, Z, \theta))|Z]]$ satisfies the existence and differentiability conditions of Definition 2.2. *Q.E.D.*

PROOF OF LEMMA A.1: If $E_\pi[\ln E_\rho[\exp(\gamma' g(U, Z, \theta))|Z]]$ exists for all $\gamma \in \mathbb{R}^{d_g}$, then $E_\rho[\exp(\gamma' g(U, Z, \theta))|Z = z]$ must exist and be finite for all $\gamma \in \mathbb{R}^{d_g}$ and for almost all z, except perhaps on a set of probability 0 under π. By the properties of moment generating functions defined for all $\gamma \in \mathbb{R}^{d_g}$, the $\frac{\partial}{\partial\gamma}$ and $\frac{\partial^2}{\partial\gamma\,\partial\gamma'}$ operators therefore commute with $E_\rho[\cdot|Z = z]$ and we have

$$\begin{aligned}
&\frac{\partial^2}{\partial\gamma_j^2} \ln E_\rho\big[\exp\big(\gamma' g(U, Z, \theta)\big)|Z = z\big] \\
&\quad= \frac{\displaystyle\int (g_j(u, z, \theta) - \tilde{g}_j(z, \theta, \gamma))^2 \exp(\gamma' g(u, z, \theta))\,d\rho(u|z)}{\displaystyle\int \exp(\gamma' g(u, z, \theta))\,d\rho(u|z)} \equiv A_j(z)
\end{aligned}$$

for $j = 1, \ldots, d_g$. Since this quantity is nonnegative at any z, we also have

$$\begin{aligned}
E_\pi\big[A_j(Z)\big] &= E_\pi\left[\frac{\partial^2}{\partial\gamma_j^2} \ln E_\rho\big[\exp\big(\gamma' g(U, Z, \theta)\big)|Z\big]\right] \\
&= \frac{\partial^2}{\partial\gamma_j^2} E_\pi\big[\ln E_\rho\big[\exp\big(\gamma' g(U, Z, \theta)\big)|Z\big]\big],
\end{aligned}$$

where the latter quantity is finite by assumption. Hence, $E_\pi[\ln E_\rho[\exp(\gamma' g(U, Z, \theta))|Z]]$ being twice differentiable implies that $A_j(Z)$ has finite expectation under π. As covariances and means can be bounded in terms of variances, the first derivatives and mixed second derivatives of $\ln E_\rho[\exp(\gamma' g(U, Z, \theta))|Z]$ also commute with the expectation $E_\rho[\cdot|Z = z]$. This in turn implies that both

$$E_\pi\left[\left|\frac{\partial}{\partial\gamma_j}\ln E_\rho\left[\exp\left(\gamma' g(U, Z, \theta)\right)|Z = z\right]\right|\right]$$

and

$$E_\pi\left[\left|\frac{\partial^2}{\partial\gamma_j\,\partial\gamma_{j'}}\ln E_\rho\left[\exp\left(\gamma' g(U, Z, \theta)\right)|Z = z\right]\right|\right]$$

are finite, and this absolute integrability result implies that $\partial/\partial\gamma_j$ and $\partial^2/\partial\gamma_j\,\partial\gamma_{j'}$ also commutes with E_π. Since we have shown that interchanges of derivatives and expectations are allowed, we can verify that

$$g_\gamma = \frac{\partial}{\partial\gamma}E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g(U, Z, \theta)\right)|Z\right]\right]$$

and

$$V_\gamma = \frac{\partial^2}{\partial\gamma\,\partial\gamma'}E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g(U, Z, \theta)\right)|Z\right]\right],$$

which both exist because $E_\pi[\ln E_\rho[\exp(\gamma' g(U, Z, \theta))|Z]]$ is twice differentiable.

To show that V_γ^{-1} exists for all $\gamma \in \mathbb{R}^{d_g}$, we show that $\eta' V_\gamma \eta$ never vanishes for any unit vector η. Note that $\eta' V_\gamma \eta$ is the expected value (under π) of the variance of $\eta' g(U, z, \theta)$ (conditional on z) under the measure $\tilde{\rho}(u|z)$ defined via

$$d\tilde{\rho}(u|z) = \exp\left(\gamma' g(u, z, \theta)\right) d\rho(u|z) \Big/ \int \exp\left(\gamma' g(u, z, \theta)\right) d\rho(u|z).$$

By assumption, $\eta' g(u, z, \theta)$ does not remain constant as u varies in $\mathcal{U}$ (for all z in a subset of positive probability under π). Since $\rho(u|z)$ is supported on all of $\mathcal{U}$ and $\exp(\gamma' g(u, z, \theta))$ is strictly positive for finite γ, it follows that the measure $\tilde{\rho}(u|z)$ is also supported on all of $\mathcal{U}$. Hence, the variance of $\eta' g(U, z, \theta)$ under $\tilde{\rho}(u|z)$ is strictly positive for any unit vector η. *Q.E.D.*

PROPOSITION B.1: *Let X and Y be random vectors* (*which could be functions of other random variables*). *If a conditional expectation* $E[Y|X]$ (*and its corre-*

sponding unconditional expectation $E[Y]$) *are well defined,*[1] *then the restriction* $E[Y|X]=0$ (*with probability* 1 *under the distribution of* X) *is equivalent to a countable set of unconditional moment restrictions*.

PROOF: By iterated expectation, it is trivial to show that $E[Y|X]=0$ (with probability 1 under F, the distribution of X) implies that $E[Ya(X)]=0$ for any measurable function $a(\cdot)$, in particular, a countable set of functions $a(\cdot)$.

To show the converse, we consider moments of the form $E[Ye^{\mathbf{i}\xi' X}]$ for $\xi\in\mathbb{R}$, where $\mathbf{i}=\sqrt{-1}$. First note that if $E[Y]$ is well defined, then $E[|Y|]$ must exist. By Lemma 3 in Schennach (2007), this implies that $E[Ye^{\mathbf{i}\xi' X}]$ is continuous in ξ. Hence, having $E[Ye^{\mathbf{i}\xi' X}]=0$ for all rational ξ implies that $E[Ye^{\mathbf{i}\xi' X}]=0$ for all $\xi\in\mathbb{R}$. The inverse Fourier transform of $E[Ye^{\mathbf{i}\xi' X}]=E[E[Y|X]e^{\mathbf{i}\xi' X}]$ therefore vanishes almost everywhere. Since $E[e^{\mathbf{i}\xi' X}E[Y|X]]=\int e^{\mathbf{i}\xi' X}E[Y|X]\,dF(x)$, its inverse Fourier transform is the measure defined via the differential element $E[Y|X=x]\,dF(x)$. Having this measure vanish almost everywhere is equivalent to having $E[Y|X=x]=0$ with probability 1 under F. Therefore, we have just shown that a countable set of unconditional moment restrictions[2] ($E[Ye^{\mathbf{i}\xi' X}]=0$ for all rational ξ) implies a conditional mean restriction ($E[Y|X]=0$ with probability 1). Note that the sequence of moments constructed here is not the only one possible (see Chamberlain (1987) for an alternative). *Q.E.D.*

PROPOSITION B.2: *Independence restrictions can be imposed via a countable number of moment factorization restriction of the form* (17), *that is*, *without loss of generality*, *the index t can be discrete*.

PROOF: Without loss of generality, let the random variables X and Y denote two random quantities (which could be functions of other random variables) to be required to be independent (more independent quantities can be handled similarly). By Theorem 16-B in Loève (1977), two random variables X and Y are independent if and only if

$$E\big[\exp(\mathbf{i}\xi X)\exp(\mathbf{i}\eta Y)\big]=E\big[\exp(\mathbf{i}\xi X)\big]E\big[\exp(\mathbf{i}\eta Y)\big] \tag{S.11}$$

for all $\xi,\eta\in\mathbb{R}$, where $\mathbf{i}=\sqrt{-1}$. By result 13.4-A in Loève (1977), all three expectations in (S.11) are continuous functions of ξ and η. Hence, imposing the constraint (S.11) at all rational ξ and η is sufficient to imply that (S.11) holds for all $\xi,\eta\in\mathbb{R}$. Since rationals are countable, the result is proven. Note that the sequence of moments constructed here is not the only possibility. *Q.E.D.*

[1] This entails measurability assumptions, absolute conditional moment existence, and regularity of the appropriate conditional measures.

[2] Note that rationals can be ordered in sequence: For instance, write them as n/m, picking $(n,m)\in\mathbb{Z}^2$ along a "square spiral pattern" and eliminating duplicates.

APPENDIX C: ADDITIONAL SIMULATION EXAMPLES

C.1. *Regression With Interval-Valued Data*

We now illustrate the method with our Example 1.1. To this effect, we use an i.i.d. sample of 250 observations, generated according to[3]

$$Y^* = X\theta_1 + V \quad \text{with} \quad \theta_1 = 1,$$
$$\overline{Y} = \lceil Y^* \rceil,$$
$$\underline{Y} = \lfloor Y^* \rfloor,$$

where $X \sim N(0, 1)$ and $V \sim N(0, 1/4)$. The algorithm of Section 2.3 (with empirical likelihood) was used with $R = 500$, after 50 equilibration steps.[4] As seen in Figure S.1, the set over which the objective function (solid curve) vanishes matches the conventional bounds (indicated by diamonds and calculated as in Manski and Tamer (2002)). This is verified analytically in Appendix H. (The small apparent discrepancy visible in the graph merely reflects the fact that the objective function is computed on a discrete mesh of values of θ_1. This qualification will apply to our remaining examples as well.) However, what is more

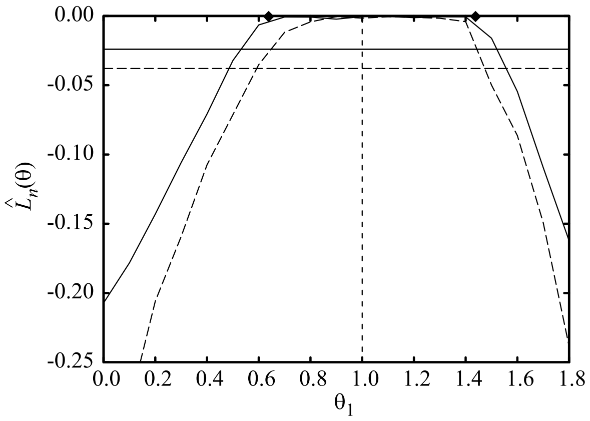

FIGURE S.1.—Objective function for an interval-valued data regression model. The upper diamonds mark the standard bounds for this model, while the true value of the parameter is indicated by a vertical dashed line. The solid curve is obtained with the usual uncorrelatedness assumption, while the dashed line is for a model that also assumes that the variance of the residuals is uncorrelated with the (squared) regressor. The horizontal solid and dashed lines indicate the corresponding critical values at the 95% level.

[3] Let $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the "round up" and "round down" operations, respectively.

[4] The number of simulation steps was determined by gradually increasing the number of steps until the simulation noise (which can be obtained by a standard variance calculation) became negligible relative to the critical value used to calculate the confidence regions.

interesting and new is that we can now easily add any other types of reasonable moment conditions we are willing to assume to narrow down the identified set (including moment conditions that may be nonmonotone in the unobservables).

EXAMPLE 1.1—Continued: The worst-case scenario that gives rise to the bounds may be associated with unusual patterns of heteroskedasticity in the residuals $Y^* - X\theta$, with point masses in the distribution of Y^* for large $|X|$ but not for small $|X|$. If this appears extremely implausible, one could add two more moment conditions to ensure that the variance of the residuals (conditional on X) is not correlated with X^2. The moment function would then be

$$g(U, Z, \theta) = \begin{bmatrix} VX \\ (V^2 - \theta_2)X^2 \\ V^2 - \theta_2 \end{bmatrix}, \tag{S.12}$$

where $V = \underline{Y} + U(\overline{Y} - \underline{Y}) - X\theta_1$ and $\theta = (\theta_1, \theta_2)$ in which θ_2 is an additional nuisance parameter (the mean of V^2).

Interestingly, this more complex model requires no additional effort on the part of the researcher—the simulations take care of everything. It would have been quite difficult to compute the bounds for this more complex model analytically, let alone to handle the sampling noise properly.

Of course, one has to take into account sampling variation so as to get a proper confidence region. This is done here by calculating a critical value and keeping all values of θ such that the objective function exceeds the critical value. Here, the critical value is obtained using Theorem F.1 (all critical values obtained in the present simulation section are obtained similarly).

C.2. *Censored Regression*

We now apply our method to the censored regression of Example 1.2 by generating an i.i.d. sample of 250 observations as

$$\begin{aligned} &X \sim N(0, 1), \\ &V \sim N(0, 1/4), \\ &Y^* = \theta_1 + X\theta_2 + V, \\ &Y = \min(Y^*, 1). \end{aligned}$$

The algorithm of Section 2.3 (with empirical likelihood) was used with $R = 900$, after 100 equilibration steps. Figure S.2(a) shows the resulting objective function. In this example, there is both an intercept and a slope parameter, but we are profiling out the intercept to only show the objective function as a

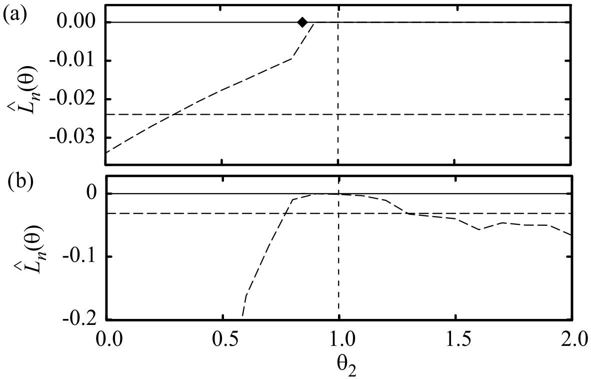

FIGURE S.2.—Objective function for a censored regression model. (a) Result obtained with the usual uncorrelatedness and zero mean assumptions on the residuals. The upper diamond marks the well known lower bound for this model. (b) Same exercise while assuming, in addition, that the variance of the residuals is uncorrelated with the regressor. In each panel, the horizontal line indicates the critical values at the 95% level and the true value of the parameter is indicated by a vertical dashed line.

function of the slope coefficient θ_2, which is of greater interest. The set over which the objective function (solid curve) vanishes matches the conventional bound (indicated by a diamond and calculated as in Manski and Tamer (2002)). Without any other information beyond the standard uncorrelatedness assumption between the regressor and the residuals, the censored regression of Example 1.2 only admits a lower bound on the slope coefficient for the (randomly generated) sample used here. No upper bound for the slope coefficient exists because the possible values of Y^* given the observed data can be arbitrarily large when there are censored observations.[5]

However, large values of the slope coefficient imply a rather strange distribution of the residuals, namely, residuals of a much larger magnitude for censored observations than for uncensored ones. By imposing slightly more structure on the residuals, it is possible to obtain both a lower and an upper bound on the slope coefficient, as shown in Figure S.2(b).

EXAMPLE 1.2—Continued: The problem of the absence of an upper bound in our censored regression example can be eliminated by simply constraining the variance of the residuals to be uncorrelated with the regressors, in addition to the usual uncorrelatedness assumption:

$$g(U, Z, \theta) = \begin{bmatrix} \big(Y + U\mathbf{1}(Y = c) - X\theta\big)X \\ \big(Y + U\mathbf{1}(Y = c) - X\theta\big)^2 X \end{bmatrix}.$$

[5] The problem still admits a lower bound because there are no censored observations below the mean of the X.

This amounts to imposing a weak form of homoskedasticity. (Note that the moment conditions here exploit the knowledge that X has zero mean, for simplicity.)

This represents a substantial reduction in the uncertainty in the model parameters. As before, this required no extra analytical work. In contrast, it would be very difficult to derive the bounds analytically because some of the moment functions are not monotone in the unobservable.

C.3. *Nonlinear Errors-in-Variables Model Without Side Information*

We now consider a model for which a preexisting analysis of identification is not available.

EXAMPLE 1.5—Continued: Consider a nonlinear errors-in-variables model

$$\begin{aligned} Y &= r(X^*, \theta) + V_2, \\ X &= X^* + V_1, \end{aligned} \tag{S.13}$$

where $r(X^*, \theta)$ is a given parametric specification with unknown parameter vector $\theta = (\theta_1, \theta_2)$ and we impose the vector of moment conditions $g(U, Z, \theta) = (V_1, V_2, V_1\,\partial r(X^*, \theta)/\partial\theta_1, V_1\,\partial r(X^*, \theta)/\partial\theta_2, V_2\,\partial r(X^*, \theta)/\partial\theta_1, V_2\,\partial r(X^*, \theta)/\partial\theta_2, V_1V_2)'$. These conditions essentially combine the uncorrelatedness assumptions of an errors-in-variables model with the standard normal equations for a least-square regression.

While it is known that this model can be point-identified under full mutual independence assumptions (Schennach and Hu (2013)), no such result exists under the weaker uncorrelatedness conditions imposed here. A sample of 250 i.i.d. observation is generated according to Equation (S.13) with $\theta_1 = 1$, $\theta_2 = 0.5$, and $X^* \sim N(0, 1)$, $V_1 \sim N(0, 1/4)$ and $V_2 \sim N(0, 1/4)$. The resulting objective function is shown in Figure S.3 for two specifications:

$$r(X^*, \theta) = \theta_1 X^* + \theta_2 (X^*)^2, \tag{S.14}$$

$$r(X^*, \theta) = \theta_1 X^* + \theta_2 \exp(X^*). \tag{S.15}$$

This example illustrates the construction of a confidence region (instead of a confidence interval). It should be noted that deriving bounds for this model would have been extremely difficult due to the nonmonotonicity of the moment functions. In fact, calculating equivalent moment inequalities from Equation (15) involves an optimization problem that has no analytic solution for the specification (S.15). In contrast, our method applies directly—only trivial changes in the program that handles the standard measurement error problem were needed.

The time needed to complete these simulations ranges from a few minutes (for the simplest models) to a few hours (for the one with 27 moment con-

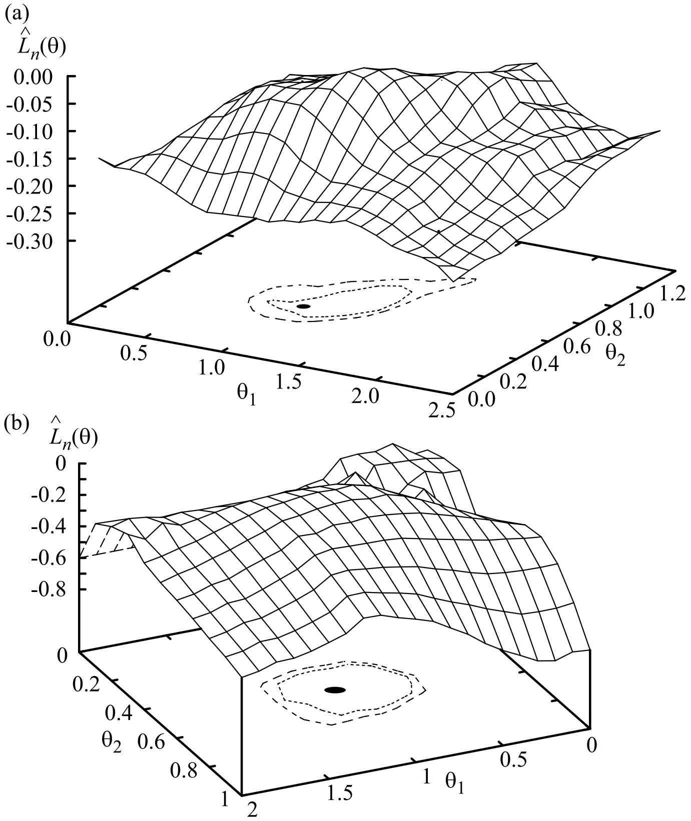

FIGURE S.3.—Objective function for (a) the polynomial measurement error model of Equation (S.14) and (b) for the linear-exponential measurement error model of Equation (S.15). The base plane shows the joint critical region at the 95% and the 99% levels, while the true value of the parameters is indicated by the filled circle.

ditions) on an average single processor personal computer in 2008–2009 and using the Gauss language. These times could undoubtedly be improved significantly by fine-tuning the implementation and using a compiled language. The main advantages of the method lie in its simplicity (regardless of the complexity of the model), its straighforward adaptability to new models, and its robustness (e.g., guaranteed convergence of the optimization algorithms thanks to smoothness and convexity).

APPENDIX D: EXTENDED NOTION OF THE IDENTIFIED SET

D.1. *Motivation*

Our extended notion of the identified set given in Equation (3) accounts for the possibility of a measure μ that does not belong to $\mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ (for instance, a distribution that is *improper* in the sense that it cannot be normalized so that $\int d\mu(u|z) = 1$ for z in a set of positive probability), but that is the limit of some sequence μ_k in $\mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ such that $E_{\mu_k \times \pi}[g(U, Z, \theta)] \to 0$. The set Θ_0 (from Equation (3)) is preferable to Θ_0^* (Equation (2)) for two reasons.

- Under Θ_0^*, the set of possible values of the moments (as $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ varies) may be open, which causes some conceptual issues in testing: Some values of the moments may, technically, be inconsistent with the model (because they are "just outside" of an open set), but there exist moment values that are arbitrarily close to those that *are* consistent with the model. This implies that any statistical test would fail to reject a model that is apparently false. These problems do not occur with Θ_0, since the set of possible values of the moments is closed by construction.
- The set Θ_0^* is not invariant to reparametrization of the dependence of $g(u, z, \theta)$ on u. An explicit example is given in Appendix D.2 below. This invariance is important because the choice of the particular parametrization of the unobservables of the model is arbitrary, as it does not result in any detectable changes in the observable quantities. In contrast, the set Θ_0 has this invariance property. This follows from the fact that the value of a supremum is the same whether the least upper bound is reached for one value of the argument or not.

The need for a more general notion of the identified set arises because we allow for moment functions $g(u, z, \theta)$, which may be unbounded or discontinuous, and sets $\mathcal{U}$, which may be unbounded. Under stronger assumptions, one can ensure $\Theta_0 = \Theta_0^*$ (e.g., Galichon and Henry (2013) make uniform integrability assumptions to rule out improper distributions). But this is unnecessary here, since, in light of the first point above, the distinction between Θ_0 and Θ_0^* is inconsequential in practice, as it could never be detected, and since the second point even emphasizes that any difference between Θ_0^* and Θ_0 would be parametrization-dependent and, therefore, meaningless. Only Θ_0 has a parametrization-independent interpretation.

D.2. *Example*

Let $g(U, Z, \theta) = \exp(-U^2) + \theta$ with $\theta \in \Theta = [0, 1]$ and U taking values in $\mathcal{U} = \mathbb{R}$. (This example does not rely on any dependence on Z; hence, the conditional distribution of U may be taken to be independent of Z without loss of generality.) We will show that Θ_0^* is empty while $\Theta_0 = \{0\}$. However, under an innocuous reparametrization of the unobservables, $\Theta_0^* = \Theta_0 = \{0\}$, thus showing that Θ_0^* is not parametrization invariant, while Θ_0 is.

Since $\sup_{\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}} E_\mu[g(U, Z, \theta)] > 0$ for all $\theta > 0$, the identified set is, at best, the singleton $\{0\}$ and we, therefore, carry out the analysis for $\theta = 0$ only.

Case 1. The case of $\mathcal{U} = \mathbb{R}$.

(a) Any proper (i.e., tight) probability measure must assign a positive probability to a compact set. Since $\exp(-U^2)$ is strictly positive on any compact nondegenerate interval, $E_\mu[\exp(-U^2)] > 0$ for any $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ and Θ_0^* is empty.

(b) However, consider a sequence of probability measure μ_j such as a sequence of Gaussians with width diverging to infinity. It can be readily verified that $E_{\mu_j}[\exp(-U^2)] \to 0$ even though $E_{\mu_j}[\exp(-U^2)] > 0$ at each j. Clearly, μ_j does not converge to a proper probability measure (the increasing width of the Gaussian causes the limit to fail to be tight). Nevertheless $\sup_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} E_\mu[\exp(-U^2)] = 0$ and we have $\Theta_0 = \{0\}$.

Case 2. Take $\tilde{U} \equiv \arctan U$ and $\tilde{g}(\tilde{U}, Z, \theta) \equiv g(\tan \tilde{U}, Z, \theta)$. By definition, the support of $\tilde{U}$ is the closure of $\{\arctan U : U \in \mathbb{R}\}$, that is, $\tilde{\mathcal{U}} = [-\pi/2, \pi/2]$. The function $\tilde{g}(\tilde{U}, Z, \theta)$ is clearly defined for $\tilde{U} \in]-\pi/2, \pi/2[$ and can naturally be extended by continuity for $\tilde{U} = \pm\pi/2$, that is, $\tilde{g}(\pm\pi/2, Z, \theta) = \lim_{\tilde{U}\to\pm\pi/2} \tilde{g}(\tilde{U}, Z, \theta) = 0 + \theta$.

(a) We then have that $\Theta_0^* = \{0\}$ because $E_\mu[\tilde{g}(\tilde{U}, Z, \theta)] = 0$ for $\theta = 0$ and μ equal to a point mass at $\tilde{U} = \pi/2$.

(b) We also have $\Theta_0 = \{0\}$ for the same reason.

Hence, in this example, Θ_0^* is not parametrization invariant, while Θ_0 is.

APPENDIX E: DIFFICULTIES WITH ALTERNATIVE DISCREPANCIES

This section shows that using likelihood maximization instead of entropy maximization leads to a solution where the Lagrange multipliers for the infinite-dimensional constraints cannot be solved for analytically.

The Lagrangian for likelihood maximization (in the notation of Section 2.2) is

$$\begin{aligned}
&-\int\int \ln\bigl(f(u|z)\bigr)\,d\rho(u|z)\,d\pi(z)\\
&\quad+\gamma'\int\int g(u,z,\theta)f(u|z)\,d\rho(u|z)\,d\pi(z)\\
&\quad+\int \phi(z)\biggl(\int f(u|z)\,d\rho(u|z)-1\biggr)\,d\pi(z).
\end{aligned}$$

The first order condition is then

$$\int\int\biggl(\frac{1}{f(u|z)} - \gamma' g(u,z,\theta) - \phi(z)\biggr)\delta f(u|z)\,d\rho(u|z)\,d\pi(z) = 0.$$

Since the equality must hold for any $\delta f(u|z)$, we have

$$\frac{1}{f(u|z)} - \gamma' g(u,z,\theta) - \phi(z) = 0$$

or, after rearranging,

$$f(u|z)=\frac{1}{\gamma' g(u,z,\theta)+\phi(z)}.$$

The fact that conditional distributions must integrate to 1 at each value of the conditioning variable implies that

$$\int \frac{1}{\gamma' g(u,z,\theta)+\phi(z)}\,d\rho(u|z)=1. \tag{S.16}$$

Clearly, $\phi(z)$ cannot be solved for analytically. Even the technique used to determine the analogue of $\phi(z)$ in conventional empirical likelihood (EL) does not work. To see this, rewrite (S.16) as

$$\begin{aligned}\left[-\int \frac{\gamma' g(u,z,\theta)}{\gamma' g(u,z,\theta)+\phi(z)}\,d\rho(u|z)\right]& \\ +\big(1-\phi(z)\big)\int \frac{1}{\gamma' g(u,z,\theta)+\phi(z)}\,d\rho(u|z)&=0.\end{aligned} \tag{S.17}$$

In EL, the first term in brackets would vanish as a consequence of the moment conditions being satisfied (thus implying that $\phi(z)$ would have to be 1). However, here, the moment conditions only imply that

$$\int\int \frac{\gamma' g(u,z,\theta)}{\gamma' g(u,z,\theta)+\phi(z)}\,d\rho(u|z)\,d\pi(z)=0$$

and the first term in (S.17) cannot be concluded to vanish (and $\phi(z)\neq 1$ in general). The distinction arises from the presence of conditional distributions in the present setup that are absent in EL.

APPENDIX F: Inference Methods

As models defined via moment conditions that involve unobservables are often set-identified, inference methods capable of handling this situation are essential. We describe below how the inferential techniques based on subsampling or other simulation techniques (as described in Chernozhukov, Hong, and Tamer (2007)) can be applied in our settings.

F.1. *Objective Functions and Confidence Regions*

We first introduce a general class of possible objective functions.

DEFINITION F.1: Given an i.i.d. sample $Z_1, \ldots, Z_n$, we consider an empirical objective function that admits the representation

$$\hat{L}_n(\theta) = \sup_{\gamma \in \mathbb{R}^{d_g}} \hat{L}_n(\theta, \gamma),$$

$$\hat{L}_n(\theta, \gamma) = -\frac{1}{2}\hat{g}'(\theta, \gamma) W(\theta, \gamma) \hat{g}(\theta, \gamma) + \hat{R}_n(\theta, \gamma),$$

where $W(\theta, \gamma)$ is a positive semidefinite[6] matrix and

$$\hat{g}(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} \tilde{g}(Z_i, \theta, \gamma),$$

and where the remainder satisfies

$$\sup_{\{\theta \in \Theta, \gamma \in \mathbb{R}^{d_g} : \|g(\theta,\gamma)\| = O(n^{-1/2})\}} \left|\hat{R}_n(\theta, \gamma)\right| = o_p\left(n^{-1}\right)$$

and is such that

$$\hat{L}_n(\theta, \gamma) \leq -C\left\|\hat{g}(\theta, \gamma)\right\|^2 \tag{S.18}$$

for some $C > 0$ with probability approaching 1. We also assume throughout that the $\rho(u|z)$ used to construct $\tilde{g}(Z, \theta, \gamma)$ is as in Definition 2.2 and that the unobservables U_i are i.i.d.

This definition includes GMM-like objective functions. In the important special case where $W(\theta, \gamma) = V^{-}(\theta, \gamma)$—the generalized inverse of $V(\theta, \gamma) = E[\tilde{g}(Z_i, \theta, \gamma)\tilde{g}'(Z_i, \theta, \gamma)]$—this definition includes the log empirical likelihood (EL) and the continuous updating estimator (CUE) as special cases:

$$\hat{L}_n^{\mathrm{EL}}(\theta) = \sup_{\gamma \in \mathbb{R}^{d_g}} \inf_{\lambda \in \mathbb{R}^{d_g}} \frac{1}{n}\sum_{i=1}^{n} -\ln\left(1 - \lambda'\tilde{g}(Z_i, \theta, \gamma)\right),$$

$$\hat{L}_n^{\mathrm{CUE}}(\theta) = \sup_{\gamma \in \mathbb{R}^{d_g}} -\frac{1}{2}\hat{g}'(\theta, \gamma)\hat{V}^{-1}(\theta, \gamma)\hat{g}(\theta, \gamma) \quad \text{with}$$

$$\hat{V}(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} \tilde{g}(Z_i, \theta, \gamma)\tilde{g}'(Z_i, \theta, \gamma).$$

[6] Even though this allows for singular weighting matrices, Equation (S.18) below prevents the objective function from vanishing when $\hat{g}(\theta, \gamma) \neq 0$.

The inclusion of EL is useful, in light of its known optimality properties in the context of point-identified models (Newey and Smith (2004), Kitamura (2001), Kitamura, Santos, and Shaikh (2012), among others) and in a large class of set-identified models (Canay (2010)). The function$\hat{L}_n(\theta)$ also includes any GEL and ETEL as special cases.

It should be noted that although $\hat{L}_n(\theta, \gamma)$ depends on the choice of ρ in Definition 2.2, the objective function $\hat{L}_n(\theta)$ does not, as can be seen by setting π to the sample distribution in Corollary 2.1.

For maximum generality, we decompose the parameter vector as $\theta = (\beta, \eta)$, where $\beta \in \mathcal{B}$ is the parameter vector of interest, while $\eta \in \mathcal{N}_\beta$ is a vector of nuisance parameters (which may be empty if desired). We focus on the construction of confidence regions for β in the identified set $\mathcal{B}_0 \equiv \{\beta \in \mathcal{B} : \inf_{\eta \in \mathcal{N}_\beta} \inf_{\gamma \in \mathbb{R}^{d_g}} \|E[\tilde{g}(Z_i, \theta, \gamma)]\| = 0\}$ via the "profiled" statistic

$$\hat{Q}_n(\beta) = -\Big(\sup_{\eta \in \mathcal{N}_\beta} \sup_{\gamma \in \mathbb{R}^{d_g}} \hat{L}_n\big((\beta, \eta), \gamma\big) - \sup_{\theta \in \Theta} \sup_{\gamma \in \mathbb{R}^{d_g}} \hat{L}_n(\theta, \gamma)\Big), \tag{S.19}$$

where $\mathcal{N}_\beta$ is a compact subset of Θ (which may be β-dependent). If no nuisance parameters are needed, the supremum over η is to be eliminated. The statistic $\hat{Q}_n(\beta)$ is positive by construction (to follow the convention of Chernozhukov, Hong, and Tamer (2007)). The idea of subtracting the maximum value of the objective function for an "unrestricted model" is known to yield efficiency improvements in point-identified models (for instance, it reduces the number of degrees of freedom of the limiting χ^2 distribution of likelihood ratio-type tests (Newey and McFadden (1994))) and it is natural to expect improvements in set-identified models. This idea is also exploited in Chernozhukov, Hong, and Tamer (2007).

In this framework, consistent estimates of the identified set and/or confidence regions have the general form

$$\hat{\mathcal{B}} = \{\beta : n\hat{Q}_n(\beta) \le \hat{c}_\alpha\}, \tag{S.20}$$

where $\hat{c}_\alpha$ is a critical value selected so that $\hat{\mathcal{B}}$ is consistent and/or has the correct coverage $1 - \alpha$.

F.2. *Consistency*

Consistency of $\hat{\mathcal{B}}$ (in the sense that the Hausdorff distance between $\hat{\mathcal{B}}$ and $\mathcal{B}_0$ goes to 0 in probability) follows by a straightforward application of Theorem 3.2 in Chernozhukov, Hong, and Tamer (2007). Most of this theorem's requisite assumptions translate directly in the present context. We focus here only on the assumptions that demand special attention.

One less obvious issue is that the set of possible values of the parameter γ is not compact (it is $\mathbb{R}^{d_g}$). This can be handled by reparametrizing the moment functions to render the parameter space compact.[7] To this effect, let $\bar{g}(\theta,\gamma)=E[\tilde{g}(Z_i,\theta,\gamma)]$, $\mathcal{K}_\theta\equiv\text{Closure}\{E_{\pi_0}[\tilde{g}(Z_i,\theta,\gamma)]:\gamma\in\mathbb{R}^{d_g}\}$, and $\mathcal{K}_\theta^*=\mathcal{K}_\theta\cap\mathcal{C}$, where $\mathcal{C}$ is a sufficiently large compact convex set that contains a neighborhood of $\{0\}$. The reparametrized moment functions are then

$$\tilde{g}_\kappa(z,\theta,\kappa)\equiv\lim_{j\to\infty}\tilde{g}(z,\theta,\gamma_j) \tag{S.21}$$

with γ_j $(j=1,2,\ldots)$ such that $\bar{g}(\theta,\gamma_j)=\kappa+(\bar{\kappa}-\kappa)/j$, where $\bar{\kappa}$ denotes the center of mass of $\mathcal{K}_\theta^*$. These definitions effectively parametrize the sample moment function by their value κ in the population. The limit in (S.21) is introduced to handle potential solutions at infinity ($\|\gamma\|\to\infty$). These solution at infinity (in γ) are mapped into solutions (in terms of κ) at the boundary of $\mathcal{K}_\theta$. Although the boundary of $\mathcal{K}_\theta$ may itself sometimes be at infinity, we can restrict $\mathcal{K}_\theta$ to a compact set $\mathcal{K}_\theta^*=\mathcal{K}_\theta\cap\mathcal{C}$ without loss of generality because we are interested in values of κ that make the sample moment as small as possible, that is, values of κ near 0. The constraint $\kappa\in\mathcal{C}$ is, therefore, not binding with probability approaching 1. The reparametrized moment functions $\tilde{g}_\kappa(\cdot,\theta,\kappa)$ are then indexed over a domain $\bigcup_{\theta\in\Theta}\{\theta\}\times\mathcal{K}_\theta^*$, which is compact by construction.

REMARK F.1: This reparametrization is merely a device in the proof of consistency—this is not needed for the implementation of the method. As explained at the end of Section 2.1, optimizing γ over a noncompact set poses absolutely no practical implementation problems. In fact, it is easier than having to worry about boundary solutions.

Another important step is to characterize the stochastic convergence of $\hat{L}_n(\theta,\gamma)$. This can be accomplished by first showing that the "tilted" moment functions $\tilde{g}(Z_i,\theta,\gamma)$ are π_0-Donsker (van der Vaart and Wellner (1996), van der Vaart (1998)), that is, their normalized sample averages converge to a tight Gaussian process in the sup metric.[8] A sufficient condition is as follows.

ASSUMPTION F.1: *The variable Z_i is i.i.d., $E[\|\tilde{g}(Z_i,\tilde{\theta},\tilde{\gamma})\|^2]<\infty$ for some $\tilde{\theta}\in\Theta$, and $\tilde{\gamma}\in\mathbb{R}^{d_g}$. For some $\alpha\in]0,1]$, the family $\tilde{g}(\cdot,\theta,\gamma)$ satisfies, for all*

[7] Of course, any regularity conditions must then apply to the reparametrized functions; otherwise, any noncompact parameter space could be made compact in this fashion without loss of generality.

[8] This result, in turn, will imply that $\hat{L}_n(\theta,\gamma)$ converges to a Gaussian process as well (over the identified set and dominated by a Gaussian process elsewhere), as a result of the permanence of the Donsker property under Lipschitz transformations.

positive δ *less than some* $\delta_0 \in]0, \infty[$,

$$\sup_{\theta_1 \in \Theta} \sup_{\gamma_1 \in \mathbb{R}^{d_g}} E\Big[\sup_{\gamma_2 \in \mathbb{R}^{d_g} : \|\bar{g}(\theta_1, \gamma_2) - \bar{g}(\theta_1, \gamma_1)\| \le \delta} \big\|\tilde{g}(Z_i, \theta_1, \gamma_2) - \tilde{g}(Z_i, \theta_1, \gamma_1)\big\|^2\Big] = O(\delta^\alpha), \tag{S.22}$$

$$\sup_{\theta_1 \in \Theta} \sup_{\gamma_1 \in \mathbb{R}^{d_g}} E\Big[\sup_{\theta_2 \in \Theta : \|\theta_2 - \theta_1\| \le \delta} \big\|\tilde{g}(Z_i, \theta_2, \gamma_1) - \tilde{g}(Z_i, \theta_1, \gamma_1)\big\|^2\Big] = O(\delta^\alpha), \tag{S.23}$$

where $\bar{g}(\theta, \gamma) = E[\tilde{g}(Z_i, \theta, \gamma)]$.

This assumption can be understood as a type of "Hölder continuity in expectation" condition. It is a very weak condition that essentially requires points of discontinuity to be rarely sampled. A violation of this assumption would involve having the boundary of the set $\mathcal{K}_\theta^*$ not be piecewise-differentiable, a somewhat pathological setting. Note that the metric used for γ in (S.22), namely $\|\bar{g}(\theta, \gamma_2) - \bar{g}(\theta, \gamma_1)\|$, ensures the Hölder condition for the reparametrized moment functions. This condition is general enough to allow for nonsmooth functions, which is important in our setting because the limit of $\tilde{g}(Z_i, \theta, \gamma)$ as $\|\gamma\| \to \infty$ may be nonsmooth in γ in the common case where the boundary of the set $\mathcal{K}_\theta$ contains "flat" portions. Allowing for nonsmooth functions is also useful to handle quantile restrictions. By Corollary 19.35 in van der Vaart (1998), Assumption F.1 implies that $\sup_{\theta \in \Theta} \sup_{\gamma \in \mathbb{R}^{d_g}} \|\tilde{g}(Z_i, \theta, \gamma)\| = O_p(n^{-1/2})$, thus providing a specific rate of uniform convergence in probability, one of the assumptions of Theorem 3.2 in Chernozhukov, Hong, and Tamer (2007). Assumption F.1 is implied by more primitive conditions on $g(u, z, \theta)$, such as moment existence and smoothness. For instance, see Lemma F.1 in Appendix F.5 for the interval-valued data model of Example 1.1.

The asymptotic treatment of Chernozhukov, Hong, and Tamer (2007) depends crucially on whether the objective functions satisfies a so-called degeneracy property. In essence, this property holds when the objective function $\hat{Q}(\beta)$ is exactly zero in a finite sample over a set that is asymptotically close to the identified set. The class of models we consider is so general that it includes objective functions that do satisfy the degeneracy property and some that do not. For instance, the interval-valued and censored data models (Examples 1.1 and 1.2) satisfy the degeneracy property, but some of the measurement error models we consider (extensions of Example 1.5 treated in Section 2.3) do not. The main implication is that regions of the type (S.20) provide root-n consistent estimates in the degenerate case (for any nonnegative constant $\hat{c}_\alpha$), but fall just short of root-n consistency (with a convergence rate of $\sqrt{\ln n/n}$) in the nondegenerate case (with $\hat{c}_\alpha \propto \ln n$).

F.3. *Critical Values*

The general subsampling techniques proposed in the context of set-identified models (Chernozhukov, Hong, and Tamer (2007) and Romano and Shaikh (2010)) can be used to obtain suitable critical values $\hat{c}_\alpha$. As noted, for example, in Imbens and Manski (2004) and Chernozhukov, Hong, and Tamer (2007), there are two main types of confidence region: pointwise regions that satisfy $\lim_{n\to\infty} P[\beta_0 \in \hat{\mathcal{B}}] \geq 1-\alpha$ for any $\beta_0 \in \mathcal{B}_0$ and "setwise" regions that satisfy $\lim_{n\to\infty} P[\mathcal{B}_0 \subset \hat{\mathcal{B}}] \geq 1-\alpha$. Each have their relative merits and domain of applicability, an issue which we will not discuss here.

In the setwise case, the critical value $\hat{c}_\alpha$ can be obtained by computing the $1-\alpha$ quantile of realizations of $\sup_{\beta\in\tilde{\mathcal{B}}} m\hat{Q}_m(\beta)$ (where $\tilde{\mathcal{B}}$ is a suitable consistent estimate of the identified set) obtained by drawing subsamples of size $m \ll n$ out of the full sample of size n.

In the pointwise case, the critical value $\hat{c}_\alpha$ is, in general, a function of β, denoted $\hat{c}_\alpha(\beta)$. It can be obtained by computing the $1-\alpha$ quantile of realizations of $m\hat{Q}_m(\beta)$ obtained by drawing subsamples of size $m \ll n$ out of the full sample of size n. An alternative critical value in the pointwise case is to set $\hat{c}_\alpha$ to be the supremum of $\hat{c}_\alpha(\beta)$ over an estimate of the identified set.[9] The latter alternative tends to produce larger regions, but avoids unsightly discontinuities in the confidence region boundary whose location unfortunately depends on user-specified parameters. As noted in Imbens and Manski (2004) and Andrews and Guggenberger (2009), it is important to ensure that pointwise confidence regions exhibit a coverage that converges uniformly (where the uniformity is with respect to the data generating process). This avoids paradoxes such as having a family of set-identified models that have a smaller confidence regions than a point-identified model nested as a special case of this family. Andrews and Guggenberger (2009) provided conditions under which pointwise regions have uniformly converging coverage in our general setup.

Most of the regularity conditions needed for the validity of subsampling invoked in Chernozhukov, Hong, and Tamer (2007) directly translate to the present setting. We focus here only on those that may require special attention. Establishing the stochastic convergence of $\hat{L}_n(\theta,\gamma)$ can be accomplished as in the consistency result (see Assumption F.1) by first showing that the tilted moment functions $\tilde{g}(Z_i,\theta,\gamma)$ are π_0-Donsker. Under some additional measurability and approximability conditions (following Chernozhukov, Hong, and Tamer (2007)), $n\hat{Q}_n(\beta)$ then admits a limiting distribution.[10]

[9] Note that this does not produce setwise coverage because the supremum of a family of quantiles is not the same as the quantile of the supremum over a family.

[10] Note that establishing that $\hat{L}(\theta,\gamma)$ is π_0-Donsker does not necessarily imply that $\hat{Q}_n(\beta)$ is (because the maximization over γ may cause loss of stochastic equicontinuity). This is an issue related to the lack of stochastic equicontinuity in moment inequality problems noted by

Another technical issue is that the set over which the maximizations take place must be sufficiently regular, that is, satisfy a condition known as Chernoff regularity (Chernoff (1954), Silvapulle and Sen (2005)). Intuitively, this requires these sets to have a boundary whose nonsmooth points consist, at worst, of kinks. This ensures that the boundary solutions in the optimization problem (which cannot be assumed away in the present settings) still result in well defined limiting distributions. While the set Θ can be directly assumed to satisfy this property, one cannot merely arbitrarily fix the set over which γ is optimized. This is handled, as for the consistency result, by reparametrizing the moment function by their expectations κ in the population. The domain of κ is $\mathcal{K}_\theta \cap \mathcal{C}$, which is a convex set because $\mathcal{K}_\theta$ is convex by construction and so is $\mathcal{C}$, by assumption. Convexity then implies Chernoff regularity (see, e.g., Claeskens (2004)).

Subsampling is not the only way to obtain critical values: one can also use the bootstrap or simulations methods that draw from the supremum of a Gaussian process (Canay (2010), Bugni (2010), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Andrews and Soares (2010), Andrews and Barwick (2012)), although the specifics of their implementation (such as the use of "shrinkage" techniques to ensure validity of the bootstrap) are not discussed here.

F.4. *Simple but Conservative Critical Values*

A working paper (Schennach (2009)) presents an asymptotic treatment that provides conservative critical values in a nearly closed form, along with a simpler computational method to construct confidence regions. This alternative treatment draws on the literature on constrained statistics methods (Silvapulle and Sen (2005), Rosen (2008)) and expresses the limiting distribution of the test statistic in terms of the so-called χ^2-bar distribution. In this fashion, a repeated optimization over γ at each resampling step is unnecessary. This method provides the limiting distribution of $\hat{L}_n(\theta)$ and confidence regions of the form

$$\Bigl\{\beta \in \mathcal{B}\colon \sup_{\eta \in \mathcal{N}_\beta} -n\hat{L}\bigl((\beta, \eta)\bigr) \leq \hat{c}_\alpha\Bigr\},$$

but does not allow for a subtraction of the objective function of the unrestricted model as in (S.19). For this reason, the resulting confidence regions tend to be conservative in general, although they are still perfectly valid. Nevertheless, in special cases, $\sup_{\theta \in \Theta} \hat{L}_n(\theta) = 0$ with probability approaching 1, and confidence regions that are not conservative can be obtained in this fashion without re-

Chernozhukov, Hong, and Tamer (2007). Nevertheless, under fairly weak conditions (see Conditions S.1 and S.3 in Chernozhukov, Hong, and Tamer (2007)), suprema over θ and γ still admit a limiting distribution.

course to resampling. This condition is closely related (though not identical) to the "degeneracy property" introduced by[11] Chernozhukov, Hong, and Tamer (2007), and is satisfied in many commonly used models, such as the interval-valued data model of Example 1.1.

We conclude this section by providing an even simpler way to calculate critical values that are also more conservative.

ASSUMPTION F.2: *The variable Z_i is i.i.d.*

ASSUMPTION F.3: *The set Θ is compact and the set $\Gamma_\theta = \{\gamma \in \mathbb{R}^{d_g} : E[\|\tilde{g}(Z, \theta, \gamma)\|] \leq C\}$ is nonempty for all $\theta \in \Theta$, for some $C < \infty$.*

ASSUMPTION F.4: *We have $E[\|\tilde{g}(Z, \theta, \gamma)\|^2] < \infty$ $\forall \theta \in \Theta$ and $\gamma \in \Gamma_\theta$.*

THEOREM F.1: *Let $\hat{L}_n(\theta)$ be as in Definition F.1 with $W(\theta, \gamma) = V^-(\theta, \gamma)$, the generalized inverse of $V(\theta, \gamma) = E[\tilde{g}(Z_i, \theta, \gamma)\tilde{g}'(Z_i, \theta, \gamma)]$. Under Assumptions F.2, F.3, and F.4, if $\theta \in \Theta_0$, then*

$$\lim_{n\to\infty} \Pr\bigl[-2n\hat{L}_n(\theta) \geq \chi^2_{d_g,\alpha}\bigr] \leq \alpha,$$

where $\chi^2_{d_g,\alpha}$ denotes the $(1-\alpha)$ quantile of the χ^2 distribution with d_g degrees of freedom $(\chi^2_{d_g})$.

PROOF: Theorem 3.4 in Owen (2001) establishes that $-2n\hat{L}_n(\theta, \gamma) \xrightarrow{d} \chi^2_q$ for $\theta \in \Theta_0$ and γ such that $E[\tilde{g}(Z, \theta, \gamma)] = 0$ with $q = \operatorname{rank}(E[\tilde{g}(Z, \theta, \gamma)\tilde{g}'(Z, \theta, \gamma)])$ for the empirical likelihood (EL) objective function. His proof first proceeds by showing that EL has the representation of Definition F.1; hence, his result applies more generally for any objective function with that representation. Note that $q \leq d_g$ and since $\chi^2_{d_g}$ stochastically dominates χ^2_q, using a $\chi^2_{d_g}$ instead of χ^2_q will produce valid, but conservative, confidence regions. It follows that $\mathcal{R} = \{(\theta, \gamma) \in \Theta \times \mathbb{R}^{d_g} : -2\hat{L}_n(\theta, \gamma) \geq \chi^2_{d_g,\alpha}\}$ is a confidence region of level $\leq \alpha$ for (θ, γ). A (slightly more) conservative region (of level $\leq \alpha$) for θ can be obtained by keeping all θ such that there exists at least one γ such that $(\theta, \gamma) \in \mathcal{R}$. This is equivalent to keeping all θ such that $\sup_{\gamma\in\mathbb{R}^{d_g}} -2n\hat{L}_n(\theta, \gamma) \geq \chi^2_{d_g,\alpha}$, that is, $-2n\hat{L}_n(\theta) \geq \chi^2_{d_g,\alpha}$. *Q.E.D.*

This theorem is useful to get a quick idea of what the confidence regions look like—a lookup in a χ^2 table is all that is needed. In some cases, the resulting region will be sufficiently small to already reject the null hypothesis of interest, in which case no further steps would be needed.

[11] Chernozhukov, Hong, and Tamer (2007) stated their degeneracy property as $\hat{L}_n(\theta) - \sup_{\theta\in\Theta} \hat{L}_n(\theta) = 0$ for all θ in a set that is asymptotically close to the true identified set.

F.5. *Primitive Conditions for Assumption F.1 in Example 1.1*

LEMMA F.1: *In Example* 1.1, *if* $E[X^4] < \infty$ *and* $E[W^4] < \infty$ (*where* $W \equiv (\overline{Y} - \underline{Y})X$), *and* W *is not degenerate at* 0, *then Assumption* F.1 *holds*.

PROOF: In this example, the moment condition is $\tilde{g}(u, z, \theta) = (\underline{y} + u(\overline{y} - \underline{y}) - x\theta)x$ with $z = (\underline{y}, \bar{y}, x)$ and we have

$$\tilde{g}(z, \theta, \gamma) = \frac{\displaystyle\int_0^1 (\underline{y} + u(\overline{y} - \underline{y}) - x\theta)x \exp(\gamma(\underline{y} + u(\overline{y} - \underline{y}) - x\theta)x)\, du}{\displaystyle\int_0^1 \exp(\gamma(\underline{y} + u(\overline{y} - \underline{y}) - x\theta)x)\, du} \tag{S.24}$$

$$= (\underline{y} - \theta x)x + \frac{w}{1 - e^{-\gamma w}} - \frac{1}{\gamma},$$

where $w \equiv (\overline{y} - \underline{y})x$. We then have, by a mean value argument, for $\theta_1 \in \Theta$ and $\gamma_1 \in \mathbb{R}^{d_g}$,

$$\begin{aligned} B_{\theta_1,\gamma_1} &\equiv E\Big[\sup_{\theta_2 \in \Theta : \|\theta_2 - \theta_1\| \le \delta} \big\|\tilde{g}(Z, \theta_2, \gamma_1) - \tilde{g}(Z, \theta_1, \gamma_1)\big\|^2\Big] \\ &\le E\Big[\sup_{\bar{\theta} \in \Theta : \|\bar{\theta} - \theta_1\| \le \delta} \big\|\nabla_{\theta'} \tilde{g}_\gamma(Z, \bar{\theta}, \gamma_1)\big\|^2\Big]\delta^2, \end{aligned}$$

where, for any $\bar{\theta} \in \Theta$ and $\gamma \in \mathbb{R}^{d_g}$,

$$\nabla_{\theta'} \tilde{g}_\gamma(Z, \bar{\theta}, \gamma) = -X^2.$$

So $\sup_{\theta_1 \in \Theta} \sup_{\gamma_1 \in \mathbb{R}^{d_g}} B_{\theta_1,\gamma_1} \le E[X^4]\delta^2 = O(\delta^2)$ if $E[X^4] < \infty$. This establishes (S.23) in Assumption F.1.

To establish (S.22), let us first relate the original and reparametrized moment functions (via $\kappa_j \equiv \bar{g}(\theta_1, \gamma_j)$ for $j = 1, 2$):

$$\begin{aligned} A_{\theta_1,\gamma_1} &\equiv E\Big[\sup_{\gamma_2 \in \mathbb{R}^{d_g} : \|\bar{g}(\theta_1,\gamma_2) - \bar{g}(\theta_1,\gamma_1)\| \le \delta} \big\|\tilde{g}(Z, \theta_1, \gamma_2) - \tilde{g}(Z, \theta_1, \gamma_1)\big\|^2\Big] \\ &= E\Big[\sup_{\kappa_2 \in \mathcal{K}_\theta : \|\kappa_2 - \kappa_1\| \le \delta} \big\|\tilde{g}_\kappa(Z, \theta, \kappa_2) - \tilde{g}_\kappa(Z, \theta, \kappa_1)\big\|^2\Big]. \end{aligned}$$

By a mean value argument, we have

$$A_{\theta_1,\gamma_1} \le E\Big[\sup_{\bar{\kappa} \in \mathcal{K}_\theta : \|\bar{\kappa} - \kappa_1\| \le \delta} \big\|\nabla_{\kappa'} \tilde{g}_\kappa(Z, \theta_1, \bar{\kappa})\big\|^2\Big]\delta^2. \tag{S.25}$$

To calculate $\nabla_{\kappa'}\tilde{g}_\kappa(Z, \theta_1, \bar{\kappa})$, we note that $\nabla_{\gamma'}\tilde{g}(z, \theta, \gamma) = \nabla_{\kappa'}\tilde{g}_\kappa(z, \theta, \kappa)\frac{\partial\kappa}{\partial\gamma'} = \nabla_{\kappa'}\tilde{g}_\kappa(z, \theta, \kappa)\nabla_{\gamma'}\bar{g}(\theta, \gamma)$, so that

$$\nabla_{\kappa'}\tilde{g}_\kappa(z, \theta, \kappa) = \nabla_{\gamma'}\tilde{g}(z, \theta, \gamma)\big(\nabla_{\gamma'}\bar{g}(\theta, \gamma)\big)^{-1} \tag{S.26}$$

for κ and γ such that $\kappa = \bar{g}(\theta, \gamma)$. So as to bound (S.26), we will now find a lower bound on $\nabla_{\gamma'}\bar{g}(\theta, \gamma)$ and then an upper bound on $\nabla_{\gamma'}\tilde{g}(z, \theta, \gamma)$. To calculate these derivatives, we note that, from (S.24), we have

$$\nabla_\gamma\tilde{g}(z, \theta, \gamma) = \nabla_\gamma\left(\frac{w}{1-e^{-\gamma w}} - \frac{1}{\gamma}\right) = \frac{1}{\gamma^2} - \frac{w^2}{(e^{\gamma w/2} - e^{-\gamma w/2})^2}, \tag{S.27}$$

where $w \equiv (\bar{y} - \underline{y})x$. Using the inequality $(e^{v/2} - e^{-v/2})^2 \geq v^2 + v^4/12$ for any $v \in \mathbb{R}$ (obtained by a Taylor expansion combined with a convexity argument), Equation (S.27) can be bounded below:

$$\frac{1}{\gamma^2} - \frac{w^2}{(e^{\gamma w/2} - e^{-\gamma w/2})^2} \geq \frac{1}{\gamma^2} - \frac{w^2}{\gamma^2 w^2 + \gamma^4 w^4/12} = \frac{1}{12/w^2 + \gamma^2}.$$

Next, we observe that if $W = (\overline{Y} - \underline{Y})X$ is not degenerate at 0, there exists $\varepsilon_1 > 0$ so that $\int_{|w|\geq\varepsilon_1} dF(w) = 1 - \varepsilon_2$ for some $\varepsilon_2 \in]0, 1[$. We can the write, after noting that the integrand is positive and increasing in w^2,

$$\begin{aligned}
&E\big[\nabla_\gamma\tilde{g}(Z, \theta, \gamma)\big] \\
&\quad\geq E\left[\frac{1}{12/W^2 + \gamma^2}\right] = \int \frac{1}{12/w^2 + \gamma^2}\,dF(w) \\
&\quad\geq \int_{|w|\geq\varepsilon_1} \frac{1}{12/w^2 + \gamma^2}\,dF(w) \geq \int_{|w|\geq\varepsilon_1} \frac{1}{12/\varepsilon_1^2 + \gamma^2}\,dF(w) \\
&\quad= \frac{1}{12/\varepsilon_1^2 + \gamma^2} \int_{|w|\geq\varepsilon_1} dF(w) = \frac{1 - \varepsilon_2}{12/\varepsilon_1^2 + \gamma^2}.
\end{aligned} \tag{S.28}$$

We now turn to the problem of finding an upper bound on $\nabla_{\gamma'}\tilde{g}_\kappa(z, \theta, \kappa)$. From (S.27)

$$\nabla_{\gamma'}\tilde{g}_\kappa(z, \theta, \kappa) = \frac{1}{\gamma^2}\left(1 - \frac{\gamma^2 w^2}{(e^{\gamma w/2} - e^{-\gamma w/2})^2}\right),$$

where one can show that $1 - \frac{v^2}{(e^{v/2} - e^{-v/2})^2} \leq v^2/12$ for any $v \in \mathbb{R}$ (by a Taylor expansion combined with a concavity argument). We can also show that $1 -$

$\frac{v^2}{(e^{v/2}-e^{-v/2})^2}$ is increasing in $|v|$ and reaches its maximum value of 1 as $|v| \to \infty$. Hence, we have $1 - \frac{v^2}{(e^{v/2}-e^{-v/2})^2} \leq \min\{v^2/12, 1\}$ and

$$\nabla_{\gamma'} \tilde{g}_\kappa(z, \theta, \kappa) \leq \frac{1}{\gamma^2} \min\{\gamma^2 w^2/12, 1\} = \min\left\{\frac{w^2}{12}, \frac{1}{\gamma^2}\right\}. \tag{S.29}$$

Combining (S.25), (S.26), (S.28), and (S.29), we have, for $\bar{\gamma}$ such that $\bar{\kappa} = \bar{g}(\theta, \bar{\gamma})$,

$$A_{\theta_1,\gamma_1} \leq \delta^2 E\left[\sup_{\bar{\kappa}\in\mathcal{K}_\theta:\|\bar{\kappa}-\kappa_1\|\leq\delta} \left(\frac{12/\varepsilon_1^2 + \bar{\gamma}^2}{1-\varepsilon_2} \min\left\{W^2/12, \frac{1}{\bar{\gamma}^2}\right\}\right)^2\right]. \tag{S.30}$$

Let $\varepsilon_3 > 0$ and consider two complementary cases.

(i) For $\kappa_1 \equiv \bar{g}(\theta, \gamma_1)$ such that $\|\kappa_1 - g(\theta, \bar{\gamma})\| \leq \delta$ for some $|\bar{\gamma}| \leq \varepsilon_3$, we use the fact that $\min(a, b) \leq a$ to write (S.30) as

$$\begin{aligned} A_{\theta_1,\gamma_1} &\leq \delta^2 E\left[\sup_{\bar{\kappa}\in\mathcal{K}_\theta:\|\bar{\kappa}-\kappa_1\|\leq\delta} \left(\frac{12/\varepsilon_1^2 + \bar{\gamma}^2}{1-\varepsilon_2} W^2/12\right)^2\right] \\ &= \delta^2 E\left[\left(\frac{12/\varepsilon_1^2 + \varepsilon_3^2}{1-\varepsilon_2} W^2/12\right)^2\right] \\ &= \delta^2 \left(\frac{12/\varepsilon_1^2 + \varepsilon_3^2}{12(1-\varepsilon_2)}\right)^2 E[W^4] \equiv \delta^2 C_1 E[W^4]. \end{aligned} \tag{S.31}$$

(ii) For all other κ_1 (those associated with $|\bar{\gamma}| > \varepsilon_3$), we use the fact that $\min(a, b) \leq b$ to write (S.30) as

$$\begin{aligned} A_{\theta_1,\gamma_1} &\leq \delta^2 E\left[\sup_{\bar{\kappa}\in\mathcal{K}_\theta:\|\bar{\kappa}-\kappa_1\|\leq\delta} \left(\frac{12/\varepsilon_1^2 + \bar{\gamma}^2}{1-\varepsilon_2} \frac{1}{\bar{\gamma}^2}\right)^2\right] \\ &= \delta^2 E\left[\sup_{\bar{\kappa}\in\mathcal{K}_\theta:\|\bar{\kappa}-\kappa_1\|\leq\delta} \left(\frac{12/\varepsilon_1^2}{1-\varepsilon_2} \frac{1}{\bar{\gamma}^2} + \frac{1}{1-\varepsilon_2}\right)^2\right] \\ &\leq \delta^2 E\left[\sup_{\bar{\kappa}\in\mathcal{K}_\theta:\|\bar{\kappa}-\kappa_1\|\leq\delta} \left(\frac{12/\varepsilon_1^2}{1-\varepsilon_2} \frac{1}{\varepsilon_3^2} + \frac{1}{1-\varepsilon_2}\right)^2\right] \\ &= \delta^2 \left(\frac{12/\varepsilon_1^2}{1-\varepsilon_2} \frac{1}{\varepsilon_3^2} + \frac{1}{1-\varepsilon_2}\right)^2 \equiv \delta^2 C_2. \end{aligned} \tag{S.32}$$

Combining (S.31) and (S.32), we have that

$$\sup_{\theta_1\in\Theta} \sup_{\gamma_1\in\mathbb{R}^{d_g}} A_{\theta_1,\gamma_1} \leq \delta^2 \max\{C_1 E[W^4], C_2\},$$

which is $O(\delta^2)$ if $E[W^4] < \infty$. This establishes (S.22) in Assumption F.1. *Q.E.D.*

APPENDIX G: COMPUTATIONAL DETAILS

Given the very different properties of the optimization problems in θ and γ, we do not jointly optimize the objective function over θ and γ. The optimization over θ is "difficult" in the sense that (i) the maximum could be reached over a set instead of at a single point (since we allow for set-identified models) and (ii) as in any nonlinear model (such as GMM), the optimization problem may have multiple local optima. In contrast, the problem of finding γ can be cast as a convex optimization problem with a unique global optimum. For these reasons, we scan over a grid of values of θ to map out the identified set and avoid any trapping in local minima. For each θ, the optimization over γ is well behaved and we use the simplex method due to Nelder and Mead (1965). This method is computationally convenient because it does not require the calculation of the derivatives of the objective function. Faster convergence of the numerical optimization could be achieved by exploiting derivatives of the objective function via a *guarded* Newton method (see Boyd and Vandenberghe (2004, Chapter 9.5.2)) or quasi-Newton method, such as the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BGFS) method (Nocedal (1980)).

APPENDIX H: EXAMPLE OF EQUIVALENCE TO ANALYTIC BOUNDS

In this section, we directly show equivalence between our approach with known analytic bounds in the simple case of Example 1.1. This verification is redundant (because we have already formally shown in Theorem 2.2 that our method correctly determines the identified set), but some readers may find this independent verification helpful.

To show this equivalence, we use the moment bounds provided by Theorem 2.2 (which is itself equivalent to the result of Theorem 2.1). In this example, $g(u, z, \theta) = (\underline{y} + u(\overline{y} - \underline{y}) - \theta x)x$ with $z = (x, \overline{y}, \underline{y})$ and $u \in \mathcal{U} = [0, 1]$. Since the unobservable is one-dimensional, the unit vector η (in Theorem 2.2) can only be $+1$ or -1.

(i) We can calculate $\lim_{r\to\infty} \eta'\tilde{g}(z, \theta, \eta r)$ for $\eta = \pm 1$,

$$\eta'\tilde{g}(z, \theta, \eta r) = \frac{\displaystyle\int_0^1 \eta g(u, z, \theta) \exp(r\eta g(u, z, \theta))\, du}{\displaystyle\int_0^1 \exp(r\eta g(u, z, \theta))\, du}$$

$$= \eta(\underline{y} - \theta x)x + \left[\eta b \frac{\left(1 + \dfrac{1}{r\eta b}(e^{-r\eta b} - 1)\right)}{1 - e^{-r\eta b}}\right]_{b=(\overline{y}-\underline{y})x},$$

and, therefore,

$$\lim_{r\to\infty} \eta'\tilde{g}(z,\theta,\eta r) = \eta(\overline{y}-\theta x)x \quad \text{if} \quad \eta x \geq 0,$$
$$\lim_{r\to\infty} \eta'\tilde{g}(z,\theta,\eta r) = \eta(\underline{y}-\theta x)x \quad \text{if} \quad \eta x < 0.$$

(ii) Equivalently, we can calculate $\sup_{u\in\mathcal{U}} \eta' g(u,z,\theta)$. If $\eta = 1$, then

$$\begin{aligned}\sup_{u\in\mathcal{U}} \eta' g(u,z,\theta) &= \sup_{u\in[0,1]} \big(\underline{y} + u(\overline{y}-\underline{y}) - \theta x\big)x \\ &= \begin{cases} (\overline{y}-\theta x)x & \text{if } x \geq 0, \\ (\underline{y}-\theta x)x & \text{if } x < 0. \end{cases}\end{aligned}$$

For $\eta = -1$, we have

$$\begin{aligned}\sup_{u\in\mathcal{U}} \eta' g(u,z,\theta) &= \sup_{u\in[0,1]} \big(\underline{y} + u(\overline{y}-\underline{y}) - \theta x\big)x \\ &= \begin{cases} -(\underline{y}-\theta x)x & \text{if } x \geq 0, \\ -(\overline{y}-\theta x)x & \text{if } x < 0. \end{cases}\end{aligned}$$

Through either route (i) or (ii), we therefore obtain the same moment inequalities:

$$(+1)E\left[\begin{Bmatrix} (\overline{y}-\theta x)x & \text{if } x \geq 0 \\ (\underline{y}-\theta x)x & \text{if } x < 0 \end{Bmatrix}\right] \geq 0,$$
$$(-1)E\left[\begin{Bmatrix} (\underline{y}-\theta x)x & \text{if } x \geq 0 \\ (\overline{y}-\theta x)x & \text{if } x < 0 \end{Bmatrix}\right] \geq 0.$$

Isolating θ yields

$$\big(E[x^2]\big)^{-1} E\left[\begin{Bmatrix} \underline{y}x & \text{if } x \geq 0 \\ \overline{y}x & \text{if } x < 0 \end{Bmatrix}\right] \leq \theta \leq \big(E[x^2]\big)^{-1} E\left[\begin{Bmatrix} \overline{y}x & \text{if } x \geq 0 \\ \underline{y}x & \text{if } x < 0 \end{Bmatrix}\right],$$

which is in agreement with, for example, Manski and Tamer (2002). The above treatment holds whether the expectation is under the population or the sample distribution, that is, it also ensures agreement in finite samples.

APPENDIX I: COMPARISON WITH OTHER METHODS

Our work has some connections with some previously proposed information-theoretic methods: Shen, Shi, and Wong (1999) suggested the use of an empirical likelihood-type objective function in the presence of unobservable variables. Their method consists of creating a discrete grid of points that approximates the support of the unobservables for each observed data point and maximizing the empirical likelihood calculated from this augmented sample, which consists of both actual data points and the created grid points. This approach has been shown to identify the true parameter value in a special case where the unobservable has a binary support. However, such a proof cannot be generalized further, because it can be verified that this method does not recover the well known bounds in the interval data model of Example 1.1.

EXAMPLE I.1: Applying the method of Shen, Shi, and Wong (1999) to Example 1.1 does not yield the correct identified set. In their method, one would create a grid of fictitious observation points within the sets $[\underline{Y}_i, \overline{Y}_i] \times X_i$. The empirical likelihood of all fictitious observation points is maximized when all points receive the same weights. The value of the slope coefficient θ_1 that corresponds to these weights is simply the slope of the regression of $(\underline{Y}_i + \overline{Y}_i)/2$ on X_i, because the uniform weights simply result in averaging values in the interval $[\underline{Y}_i, \overline{Y}_i]$. Now, if instead one places all the weight on $\overline{Y}_i$ for $X_i > 0$ and all weight on $\underline{Y}_i$ for $X_i < 0$, the corresponding θ_1 parameter is the slope of the regression of $\overline{Y}_i \mathbf{1}[X_i > 0] + \underline{Y}_i \mathbf{1}[X_i < 0]$ on X_i. This is, in general, a different value of θ_1 that is nevertheless equally plausible (one cannot rule out that the dependent variable takes these specific values). Yet, the value of the empirical likelihood for this set of weights is much lower (in fact, it is 0). Hence, the method assigns a different likelihood to two equally likely values of the slope parameter θ_1.

Our proposed method may be reminiscent of various entropy maximization methods proposed in Golan, Judge, and Miller (1996). Like Shen et al.'s method, discretization of the unobservables is built into the method and its computational requirements scale rapidly with the number of created support points for the unobservables. A crucial distinction with our method is the fact that their method is aimed at problems where the unobservables are variables such as the disturbances in a conventional least-square regression (note that their method does not reduce to conventional least-squares in such a case). Genuinely unobservable variables, as considered here, are not investigated in Golan, Judge, and Miller (1996) and subsequent work.

REFERENCES

ANDREWS, D. W. K., AND P. GUGGENBERGER (2009): "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709. [22]

ANDREWS, D. W. K., AND P. J. BARWICK (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826. [23]

ANDREWS, D. W. K., AND G. SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [23]

BOYD, S., AND L. VANDENBERGHE (2004): *Convex Optimization*. New York: Cambridge University Press. [28]

BUGNI, F. (2010): "Bootstrap Inference in Partially Identified Models," *Econometrica*, 78, 735–753. [23]

CANAY, I. (2010): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," *Journal of Econometrics*, 156, 408–425. [19,23]

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334. [9]

CHERNOFF, H. (1954): "On the Distribution of the Likelihood Ratio," *The Annals of Mathematical Statistics*, 25, 573–578. [23]

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284. [17,19,21-24]

CLAESKENS, G. (2004): "Restricted Likelihood Ratio Lack-of-Fit Tests Using Mixed Spline Models," *Journal of the Royal Statistical Society, Ser. B*, 66, 909–926. [23]

GALICHON, A., AND M. HENRY (2013): "Dilation Bootstrap," *Journal of Econometrics*, 177, 109–115. [15]

GOLAN, A., G. JUDGE, AND D. MILLER (1996): *Maximum Entropy Econometrics: Robust Estimation With Limited Data*. New York: Wiley. [30]

IMBENS, G., AND C. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857. [22]

KITAMURA, Y. (2001): "Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 69, 1661–1672. [19]

KITAMURA, Y., A. SANTOS, AND A. M. SHAIKH (2012): "On the Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 80, 413–423. [19]

LOÈVE, M. (1977): *Probability Theory I*. New York: Springer. [9]

MANSKI, C. F., AND E. TAMER (2002): "Inference on Regressions With Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546. [10,12,29]

NELDER, J., AND R. MEAD (1965): "A Simplex Method for Function Minimization," *Computer Journal*, 7, 308–313. [28]

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engel and D. L. McFadden. Amsterdam: Elsevier. [19]

NEWEY, W., AND R. J. SMITH (2004): "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255. [19]

NOCEDAL, J. (1980): "Updating Quasi-Newton Matrices With Limited Storage," *Mathematics of Computation*, 35, 773–782. [28]

OWEN, A. B. (2001): *Empirical Likelihood*. New York: Chapman & Hall/CRC. [24]

ROMANO, J. P., AND A. M. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211. [22,23]

ROSEN, A. M. (2008): "Confidence Sets for Partially Identified Parameters That Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107–117. [23]

SCHENNACH, S. M. (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201–239. [9]

——— (2009): "Simple Conservative Confidence Regions for a Class of Set Identified Models," Working Paper, University of Chicago. [23]

SCHENNACH, S. M., AND Y. HU (2013): "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information," *Journal of the American Statistical Association*, 108, 177–186. [13]

SHEN, X., J. SHI, AND W. H. WONG (1999): "Random Sieve Likelihood and General Regression Models," *Journal of the American Statistical Association*, 94, 835–846. [29,30]

SILVAPULLE, M. J., AND P. L. SEN (2005): *Constrained Statistical Inference*. Hoboken, NJ: Wiley. [23]

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge: Cambridge University Press. [20, 21]

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes. With Applications to Statistics*. New York: Springer. [20]

Dept. of Economics, Brown University, Providence, RI 02912, U.S.A.; smschenn@brown.edu.

Manuscript received November, 2009; final revision received June, 2013.