Schennach, Susanne

Working Paper

Entropic latent variable integration via simulation

cemmap working paper, No. CWP32/13

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Schennach, Susanne (2013) : Entropic latent variable integration via simulation, cemmap working paper, No. CWP32/13, Centre for Microdata Methods and Practice (cemmap), London,
https://doi.org/10.1920/wp.cem.2013.3213

This Version is available at:
https://hdl.handle.net/10419/79553

cemmap
centre for microdata methods and practice

Entropic latent variable integration via simulation

Susanne Schennach

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP32/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Entropic Latent Variable Integration via Simulation

Susanne M. Schennach*
Department of Economics
Brown University
smschenn@brown.edu

Initial submission: November 6, 2009
This version: June 24, 2013

Abstract

This paper introduces a general method to convert a model defined by moment conditions involving both observed and unobserved variables into equivalent moment conditions involving only observable variables. This task can be accomplished without introducing infinite-dimensional nuisance parameters using a least-favorable entropy-maximizing distribution. We demonstrate, through examples and simulations, that this approach covers a wide class of latent variables models, including some game-theoretic models and models with limited dependent variables, interval-valued data, errors-in-variables, or combinations thereof. Both point- and set-identified models are transparently covered. In the latter case, the method also complements the recent literature on generic set-inference methods by providing the moment conditions needed to construct a GMM-type objective function for a wide class of models. Extensions of the method that cover conditional moments, independence restrictions and some state-space models are also given.

Keywords: method of moments, latent variables, unobservables, partial identification, entropy, simulations, least-favorable family.

*The author would like to thank Daniel Wilhelm, seminar participants at numerous universities and at the 2010 World Congress of the Econometrics Society, as well as five anonymous referees and the co-editors for useful comments and acknowledges support from the National Science Foundation via grant SES-0752699 and SES-1061263/1156347.

1 Introduction

1.1 Outline

Our goal is to find the value(s) of a parameter $\theta \in \mathbb{R}^{d_\theta}$ that satisfy a set of moment conditions that are known to hold in the population. Unlike the conventional Generalized Method of Moments (GMM) (Hansen (1982)), we consider models where some of the variables entering the moment conditions are not observable. Specifically, the moment conditions have the general form:

$$E\left[g\left(U, Z, \theta\right)\right] = 0, \tag{1}$$

where g is a d_g-dimensional vector of nonlinear measurable functions depending on the parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, on an *unobserved* random vector U taking value in $\mathcal{U} \subseteq \mathbb{R}^{d_u}$ and on an observed random vector Z taking value in $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$. These moment conditions can be underidentified, just-identified or overidentified. We present a general method that covers this wide class of models while avoiding any parametric assumptions (beyond the given functional form of g) and without introducing any infinite-dimensional nuisance parameters, through the use of a low-dimensional dual representation of the identification problem. The use of a dual representation for this purpose has been previously suggested in the important works of Galichon and Henry (2006) and Ekeland, Galichon, and Henry (2010). This paper's contribution is to observe that a different dual formulation offers considerable advantages in terms of computational simplicity, conceptual interpretation (via a least-favorable family of distributions enabling a simple nonparametric generalization of the method of simulated moments), and in terms of weakening the necessary regularity conditions (notably, allowing for unbounded moment functions $g\left(u, z, \theta\right)$, such as the mean of a variable with unbounded support).

In essence, the method consists of eliminating the unobservables by averaging the moment function $g\left(U, Z, \theta\right)$ over these unobservables using a least-favorable distribution (i.e. one that does not make the estimation problem artificially easier) obtained through an entropy maximization procedure. This averaging can be conveniently carried out via simulations, hence the name "Entropic Latent Variable Integration via Simulation (ELVIS)". The result is a set of conventional moment conditions involving only observable variables that can be cast into a GMM-type objective function or any of its convenient one-step alternatives, such as Empirical Likelihood (EL) (Owen (1988)) or its generalizations (GEL, ETEL) (see, for instance, Owen (1990), Qin and Lawless (1994), Kitamura and Stutzer (1997), Imbens, Spady, and Johnson (1998),

Newey and Smith (2004), Schennach (2007b)). Although the unobservables have been "integrated out" from the original moment conditions, the resulting averaged moment conditions are formally equivalent to the original moment conditions, in the sense that the values of θ solving the averaged moment conditions are the same as the values solving the original moment conditions.

Latent variable models are often set-identified, that is, Equation (1) often admits more than one θ as a solution. The proposed method bypasses the complex task of establishing point- or set-identification of the model by providing a vector of moment conditions that are, by construction, satisfied (asymptotically) over the identified set, whatever it may be. General methods aimed at carrying out accurate statistical inference in set-identified models (where the set may be reduced to a single point) are being actively developed (e.g., Chernozhukov, Hong, and Tamer (2007), Andrews and Jia (2008), Beresteanu and Molinari (2008), Chiburis (2008), Rosen (2008), Andrews and Guggenberger (2009), Andrews and Soares (2010), Bugni (2010), Canay (2010), Romano and Shaikh (2010), Chernozhukov, Kocatulum, and Menzel (2012)). While these very general methods are applicable to a wide variety of user-specified objective functions, they provide little guidance on how to construct the objective function (e.g. via deriving suitable moment inequalities) for the general class of latent variable models we consider here. Our contribution is thus entirely complementary to this growing literature, as it provides specific feasible moment conditions that can be used to construct a GMM-type objective function that is compatible with many of these inference methods. This objective function is also compatible with traditional inference methods (e.g. Hansen (1982), Newey and McFadden (1994)) when the model happens to be known to be point-identified.

The paper is organized as follows. We first give a series of simple examples motivating the usefulness of the class of model considered. We then describe the method, both at a formal and at a more intuitive level, before comparing it with existing methods. A number of important extensions of the method are also described, enabling the treatment of conditional mean and independence restrictions as well as some state-space models. Finally, the capabilities of the proposed method are illustrated via simulations experiments. All proofs can be found in Appendix A or in Section C of the Supplementary Material. The Supplementary Material describes how existing general inference techniques (such as Chernozhukov, Hong, and Tamer (2007)) can be used to construct consistent set-estimates and suggests a simple, but conservative, alternative method based on a χ^2 approximation.

1.2 Motivating examples

A few straightforward examples are helpful to illustrate the very general class of models that can be handled. Simplifications, such as linearity or separability, are made for simplicity of exposition, but are in no way necessary.

Example 1.1 (Interval-valued data regression) *Consider the model*

$$Y^* = X\theta_1 + V,$$

where the scalar regressor X is perfectly observed but where the scalar dependent variable Y^* is not directly observable. Instead, it is known to lie in an interval $\left[\underline{Y}, \overline{Y}\right]$ which may vary across individuals. The scalar disturbance V satisfies $E[VX] = 0$. This model fits our framework with $Z = \left(\underline{Y}, \overline{Y}, X\right)'$, $U = \left(Y^* - \underline{Y}\right) / \left(\overline{Y} - \underline{Y}\right)$, $\mathcal{U} = [0,1]$ and

$$g(U, Z, \theta) = \left(\underline{Y} + U\left(\overline{Y} - \underline{Y}\right) - X\theta_1\right) X.$$

where we have normalized the unobservable variable U to be supported on $[0,1]$ for convenience.

Example 1.2 (Censored regression) *Consider the model*

$$Y^* = \theta_1 + X\theta_2 + V,$$

with $E[V] = 0$ and $E[VX] = 0$, and where the scalar regressor X is perfectly observed but where the scalar dependent variable Y^* is not directly observable.[1] *Instead, one observes*

$$Y = \min\left(Y^*, c\right)$$

for some known constant c. This model fits our framework with $Z = (Y, X)'$, $U = Y^* - Y$, $\mathcal{U} = \mathbb{R}^+$ and

$$g(U, Z, \theta) = \begin{bmatrix} \left(Y + U\mathbf{1}\left(Y = c\right) - \theta_1 - X\theta_2\right) \\ \left(Y + U\mathbf{1}\left(Y = c\right) - \theta_1 - X\theta_2\right) X \end{bmatrix},$$

letting $\mathbf{1}(\cdot)$ denote the indicator function. Of course, at best, this model only implies a one-sided bound on θ_2, but additional reasonable moment constraints can be easily added to address this, as we will later see. Although it is known that a linear censored regression model is point-identified under a conditional median assumption, it is nevertheless interesting to see the implications of maintaining the usual least-squares assumption in this context.

[1] The moment conditions in Examples 1.1 and 1.2 are selected so as to provide the best linear predictor, in the sense of Ponomareva and Tamer (2010), even in the presence of potential model misspecification.

Example 1.3 (Moment inequalities) *Consider a model defined via the vector of moment inequalities* $E[b(U_1, Z, \theta)] \geq 0$ *(where the inequality holds element by element). This model can be re-cast into our framework by defining*

$$g(U, Z, \theta) = b(U_1, Z, \theta) - U_2,$$

where $U = (U_1, U_2)$ *and* U_2 *is an unobserved vector of positive random variables. There may be no practical benefits associated with rewriting the original model as such, but this example demonstrates that the class of models considered here is a generalization of moment inequality models.*

Example 1.4 (Game theoretic and choice models) *Consider a model where an agent receives (parametrically specified) expected payoffs* $p(c, X, U, \theta)$ *if he picks choice* $c \in \mathcal{C}$*, where* X *is a vector of observed covariates,* U *is an unobservable disturbance (known to the agent but not to the econometrician) and* θ *is the parameter of interest. The econometrician observes the choice made* C*, and infers that* $p(C, X, U, \theta) \geq p(c, X, U, \theta)$ *for all* $c \in \mathcal{C} \backslash \{C\}$*. The use of such "revealed preference" argument has long history in economics (Afriat (1973), Varian (1982)) and still constitutes a very active area of empirical and theoretical investigation (Haile and Tamer (2003), Blundell, Browning, and Crawford (2005), McFadden (2005), Pakes, Porter, Ho, and Ishii (2005) and Example 3 in Chernozhukov, Hong, and Tamer (2007)). A special feature of our approach is that it allows for the disturbances* U *to enter the expected payoffs in a nonlinear, nonmonotone and nonseparable fashion. The revealed preference argument alone may not yet provide very much information regarding* θ *since it only sets the support of* U *for given* X *and* θ*. However, if a vector of instruments* W *is observed, one can include the restriction* $E[UW] = 0$ *to narrow down the identified set.*[2] *This model fits our framework with*[3] $Z = (C, X', W')'$*,* $\mathcal{U}$=$\mathbb{R}$ *and*

$$g(U, Z, \theta) = \begin{bmatrix} UW \\ 1 - \prod_{c \in \mathcal{C}} 1[p(C, X, U, \theta) \geq p(c, X, U, \theta)] \end{bmatrix}.$$

The second moment condition imposes that the fraction of the population satisfying all the necessary payoff inequalities is 1. More generally, U *could be a vector (and the function* p *could extract some of its components, based on the* c *argument). Also, multiplayer games can be handled, with payoffs of the form* $p_j(C_j, C_{-j}, X, U, \theta)$ *for*

[2] Although the identified set may not necessarily shrink down to a single point, even if, without loss of generality, some of the payoff functions are normalized to zero.

[3] Alternatively, one may eliminate the second moment condition and use a z and θ-dependent support for U, namely $\mathcal{U}_{z,\theta} \equiv \{u \in \mathbb{R} : p(c, x, u, \theta) \geq p(\tilde{c}, x, u, \theta) \text{ for all } \tilde{c} \in \mathcal{C} \backslash \{c\}\}$.

player $j \in \mathcal{J}$ taking action C_j, while his opponents take actions C_{-j}, leading to constraints on U of the form $p_j(C_j, C_{-j}, X, U, \theta) \geq p_j(c, C_{-j}, X, U, \theta)$ for all $c \in \mathcal{C} \setminus \{C_j\}$ and all $j \in \mathcal{J}$.

Example 1.5 (Errors-in-variables) *Consider a model with an observable scalar dependent variable Y, a scalar disturbance V_1, and an unobserved scalar regressor X^* whose observed counterpart, X, is measured with error V_2:*

$$\begin{aligned} Y &= X^*\theta + V_2 \\ X &= X^* + V_1. \end{aligned}$$

A natural set of moment conditions in this case could be:

$$E[V_1] = 0,\ E[V_2] = 0,\ E[X^*V_1] = 0,\ E[X^*V_2] = 0 \text{ and } E[V_1V_2] = 0.$$

Even though a dataset would not contain values of X^*, V_1 and V_2, this model effectively has only one unobservable. Without loss of generality, let us select X^* as our unobservable and note that all other variables then acquire unique values through

$$\begin{aligned} V_2 &= Y - X^*\theta \\ V_1 &= X - X^*. \end{aligned}$$

This model fits our framework with $Z = (Y, X)'$, $U = X^*$, $\mathcal{U} = \mathbb{R}$ and

$$g(U, Z, \theta) = \begin{bmatrix} (X - U) \\ (Y - U\theta) \\ U(X - U) \\ U(Y - U\theta) \\ (X - U)(Y - U\theta) \end{bmatrix}.$$

Remark 1.1 *It should be clear from the above examples that, in our framework,* unobservable *variables are those whose values are* not *uniquely determined once the observable variables and the parameters are known. For instance, the error term in conventional regression is* not *considered an unobservable variable. Similarly, the two disturbances in a conventional two-equation instrumental variable regression are not considered unobservable.*

While these examples are fairly simple, we will later see (in Section 5) how adding more moment conditions will lead to substantial reductions in the uncertainty in the model parameters. The proposed method is especially suited to such an exercise because it requires no extra analytical work, even in cases where it would be very difficult to derive the bounds analytically (e.g. when some of the moment functions are not monotone in the unobservables).

2 Method

2.1 Formal result

We first state definitions and conventions used throughout. Random variables (including random vectors) are denoted by capital letters and the corresponding lowercase letters represent specific values of these variables. All random variables taking value in some specified set (subsets of $\mathbb{R}^d$ for some d) have an associated probability space based on the corresponding Borel sigma-field. All functions are assumed measurable under that sigma-field and so are all sets.

Definition 2.1 *Let $\mathcal{P}_{\mathcal{S}}$ denote the set of all probability measures supported on the set $\mathcal{S}$ or any of its measurable subsets. Let $\mathcal{P}_{\mathcal{S}|\mathcal{C}}$ denote the set of all regular (see Dudley (2002), ch.10.2) conditional probability measures[4] supported on $\mathcal{S}$ (or any of its measurable subsets) given events that are measurable subsets of $\mathcal{C}$. For $\pi \in \mathcal{P}_{\mathcal{C}}$ and $\rho \in \mathcal{P}_{\mathcal{S}|\mathcal{C}}$, we let $\nu \equiv \rho \times \pi$ denote the measure $\nu \in \mathcal{P}_{\mathcal{S}\times\mathcal{C}}$ defined by products of conditional probabilities under ρ by probabilities under π. In an integral with respect to ν, the differential element $d\nu(s,c)$ will be written as $d\rho(s|c)\,d\pi(c)$ where $s \in \mathcal{S}$ and $c \in \mathcal{C}$. Whenever a conditional measure $\rho(\cdot|\cdot)$ depends on some parameter θ, it will be denoted $\rho(\cdot|\cdot;\theta)$. Let $E_{\mu}[\cdot]$ denote expectation with respect to the probability measure μ. If the subscript is omitted, the expectation is under the true data generating process. Let $\|\cdot\|$ denote the Euclidian norm for vectors and matrices.*

Assumption 2.1 *The marginal distribution of Z is supported on some set $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$, while the distribution of U conditional on $Z = z$ is supported on or inside[5] the set $\mathcal{U} \subseteq \mathbb{R}^{d_u}$ for any $z \in \mathcal{Z}$. The parameter vector θ belongs to a compact set $\Theta \subseteq \mathbb{R}^{d_\theta}$.*

Remark 2.1 *Supports are closed by definition, so, in particular, $\mathcal{U}$ and $\mathcal{Z}$ are closed. Without loss of generality, we suppress the dependence of the set $\mathcal{U}$ on z or θ. Such a dependence can always be eliminated by rewriting an equivalent estimation problem in*

[4] In general, the set $\mathcal{P}_{\mathcal{S}|\mathcal{C}}$ may depend on the probability measure π assigned to $\mathcal{C}$, but this is suppressed in the notation for conciseness.

[5] The qualifier "on or inside the set $\mathcal{U}$" takes into account the fact that some points θ of the identified set (e.g., boundary points) may be associated with distributions supported on a set smaller than $\mathcal{U}$. Indeed, one can construct a sequence of distributions supported on $\mathcal{U}$ (whose moments converge to some limiting value) but that converges to a distribution supported on a set smaller than $\mathcal{U}$. A typical example is the case of interval-valued data, where the boundaries of the identified set are associated with point masses in the distribution of the unobservables. This cannot always be avoided by simply reducing the size of the set $\mathcal{U}$, because different θ may correspond to distributions with different supports This is not a limitation or an artifact of the present approach, but is a necessary universal feature of moment condition models with unobservables.

which the dependence of $\mathcal{U}$ on z or θ has been incorporated into the moment function g. This is illustrated in our earlier Examples 1.1 and 1.2 and discussed in Section 4.1 below. When implementing the method, it is not necessary to transform the model onto a form where $\mathcal{U}$ does not depend on z or θ (since a z and θ-dependent support is trivial to account for). This reparametrization is done in the paper solely to simplify the notation.

Remark 2.2 *The sets $\mathcal{U}$ and $\mathcal{Z}$ need not be bounded, although it is clear that, to minimize the size of the identified set, researchers should select the set $\mathcal{U}$ to be as small as possible given the model's assumptions. In many popular models (e.g. Examples 1.1 and 1.2) the choice of $\mathcal{U}$ is obvious. Taking $\mathcal{U}$ to be larger than the actual support of U results in valid but conservative identified sets, an observation that proves useful when the choice of $\mathcal{U}$ is less obvious. If nothing is known regarding the support $\mathcal{U}$, one may set $\mathcal{U} = \mathbb{R}^{d_u}$ and this choice may still yield useful bounded identified sets. For instance, the measurement error model of Example 1.5 is a case where an unbounded set $\mathcal{U}$ yields a bounded identified set. The validity of Theorem 2.1 below is not affected by a conservative choice of the set $\mathcal{U}$, because this would affect Equations (5) and (6) in the same way.*

Let $\pi \in \mathcal{P}_Z$ denote the probability measure of the observable variables, with π_0 denoting the true probability measure of the observable variables. Traditionally, the *identified set* Θ_0^* is defined as (e.g., Roehrig (1988), Ekeland, Galichon, and Henry (2010)):[6]

$$\Theta_0^* = \left\{\theta \in \Theta : \text{there exists a } \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}} \text{ such that } E_{\mu\times\pi_0}\left[g\left(U,Z,\theta\right)\right] = 0\right\}. \quad (2)$$

Note that $\mu \times \pi_0$ is **not** necessarily equal to the true joint probability measure of U and Z, even for $\theta \in \Theta_0^*$.

In our treatment, it is natural to slightly extend the notion of the identified set in (2) as follows:

$$\Theta_0 = \left\{\theta \in \Theta : \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \left\|E_{\mu\times\pi_0}\left[g\left(U,Z,\theta\right)\right]\right\| = 0\right\}. \quad (3)$$

As discussed in Section E of the Supplementary Material, the refinement (3) avoids conceptual difficulties in testing (associated with having a potentially open set of possible values of the moments) and ensures invariance of the identified set under

[6] To simplify the notation, it is understood that a statement of the form $E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right] = 0$ means "$E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right]$ is well defined (i.e. $E_{\mu\times\pi}\left[\left\|g\left(U,Z,\theta\right)\right\|\right] < \infty$) and $E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right] = 0$.

observationally equivalent reparametrizations of the model's unobservables. 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.

Our method requires a user-specified dominating conditional measure ρ for the distribution of the unobservables given the observables. The exact choice of ρ has no effect on the results as long as it satisfies the properties below.[7] In general, ρ may be θ-dependent, hence we use the notation $\rho(\cdot|\cdot;\theta)$.

Definition 2.2 *For any $\theta \in \Theta$, let $\rho(\cdot|\cdot;\theta) \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ be such that*

1. $\operatorname{supp}\rho(\cdot|z;\theta) = \mathcal{U}$ *for each $z \in \mathcal{Z}$ and*
2. $E_\pi\left[\ln E_{\rho(\cdot|\cdot;\theta)}\left[\exp\left(\gamma' g(U,Z,\theta)\right)|Z\right]\right]$ *exists and is twice differentiable in γ for all $\gamma \in \mathbb{R}^{d_g}$.*

While measures $\rho(u|z;\theta)$ satisfying the above restrictions are easy to construct, the following proposition is useful to construct a suitable $\rho(u|z;\theta)$ automatically.

Proposition 2.1 *The $\rho(\cdot|\cdot;\theta) \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ in Definition 2.2 always exists: For instance, select $q \in]0,1[$ and $\omega \in]0,1[$, and for each $z \in \mathcal{Z}$ and $\theta \in \Theta$, select $\dot{u}(z,\theta) \in \mathcal{U}$ such that $\|g(\dot{u}(z,\theta),z,\theta)\| \leq \inf_{u\in\mathcal{U}}\|g(u,z,\theta)\| + \omega$. Then set*[8]

$$d\rho(u|z;\theta) = C(z,\theta)\exp\left(-\|g(u,z,\theta) - g(\dot{u}(z,\theta),z,\theta)\|^2\right) d\lambda(u|z;\theta), \qquad (4)$$

where $C(z,\theta) = \left(E_{\lambda(\cdot|\cdot;\theta)}\left[\exp\left(-\|g(U,z,\theta) - g(\dot{u}(z,\theta),z,\theta)\|^2\right)|Z=z\right]\right)^{-1}$ is a normalization constant and $\lambda(\cdot|\cdot;\theta)$ is a conditional probability measure satisfying $\operatorname{supp}\lambda(\cdot|z;\theta) = \mathcal{U}$ *and that has a point mass of probability q at $\dot{u}(z,\theta)$ conditional on $Z = z$.*

Although the above choice of ρ provides a way to secure a universal result, in almost all reasonable (and practically useful) cases, a considerably simpler choice is equally valid. For instance, the centering by $g(\dot{u}(z,\theta),z,\theta)$ is often unnecessary (e.g. when $\inf_{u\in\mathcal{U}}\|g(u,z,\theta)\|$ is zero or uniformly bounded in z and θ). The point mass

[7] We view this as a definition rather than an assumption, since ρ can be chosen (unlike the data generating process).

[8] A statement of the form $d\rho(u|z) = a(u,z)\,d\lambda(u)$ for some function $a(u,z)$, normalized so that $\int a(u,z)\,d\lambda(u) = 1$ for any z, is to be understood as "$a(\cdot,z)$ is the Radon-Nikodym derivative of $\rho(\cdot|z)$ with respect to λ, i.e. $d\rho(\cdot|z)/d\lambda = a(\cdot,z)$."

in λ is also usually not needed (e.g. whenever $\dot{u}(z,\theta)$ can be chosen such that the point $g(\dot{u}(z,\theta),z,\theta)$ remains sufficiently far from the boundary of the convex hull of $\{g(u,z,\theta) : u \in \mathcal{U}\}$). Moreover, ρ that are θ-independent are typically possible if standard dominance conditions on $g(u,z,\theta)$ hold. We are now ready to state our main identification result and a convenient corollary (both proven in Appendix A).

Theorem 2.1 *Let Assumption 2.1 hold. For any $\theta \in \Theta$ and $\pi \in \mathcal{P}_{\mathcal{Z}}$,*

$$\inf_{\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \|E_{\mu\times\pi}[g(U,Z,\theta)]\| = 0 \tag{5}$$

if and only if[9]

$$\inf_{\gamma \in \mathbb{R}^{d_g}} \|E_{\pi}[\tilde{g}(Z,\theta,\gamma)]\| = 0, \tag{6}$$

where

$$\tilde{g}(z,\theta,\gamma) \equiv \frac{\int g(u,z,\theta)\exp(\gamma' g(u,z,\theta))\, d\rho(u|z;\theta)}{\int \exp(\gamma' g(u,z,\theta))\, d\rho(u|z;\theta)}, \tag{7}$$

where $\rho(\cdot|\cdot;\theta) \in P_{\mathcal{U}|\mathcal{Z}}$ is the user-specified conditional probability measure from Definition 2.2.

Corollary 2.1 *Under Assumption 2.1 and for $\rho(\cdot|\cdot;\theta) \in P_{\mathcal{U}|\mathcal{Z}}$ as in Definition 2.2, for any $\theta \in \Theta$ and $\pi \in \mathcal{P}_{\mathcal{Z}}$,*

$$\textit{Closure}\left\{E_{\mu\times\pi}[g(U,Z,\theta)] : \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}\right\} = \textit{Closure}\left\{E_{\pi}[\tilde{g}(Z,\theta,\gamma)] : \gamma \in \mathbb{R}^{d_g}\right\}.$$

Theorem 2.1 proves that the infinite-dimensional problem of establishing the existence of some measure μ solving the original moment condition is equivalent to the much simpler problem of establishing that a finite-dimensional parameter γ solves a modified moment condition (6). This is not only convenient, but opens the way to simple estimation methodologies that are free of bias-variance trade-offs. It improves upon the intuitive approach of substituting a series approximation to the distribution of the unobservables into the method of simulated moments (McFadden (1989), Pakes and Pollard (1989)). In such an approach, the truncation of the series would result in a bias that is absent in our method.

The modified moment condition involves a function $\tilde{g}(z,\theta,\gamma)$ that is just an average of the original moment condition $g(U,z,\theta)$ under some distribution of the unobservables that belongs to a specific exponential family. Corollary 2.1 tells us that

[9] The norm in equation (6) need not be the Euclidian norm, thanks to the equivalence of all norms in finite-dimensional spaces. For instance, one could use a reciprocal-variance-weighted Euclidian norm, in analogy with efficient GMM.

what is special about the exponential family selected is its "least-favorable" property, i.e., it can reproduce the same range of values of the expectation of $g(U, Z, \theta)$ as the set of every possible conditional distribution supported on $\mathcal{U}$ given Z. In addition to the general proof of this equivalence found in Appendix A, Section I in the Supplementary Material provides, as an example, an explicit verification that our approach matches existing bounding results in the well-known special case of interval-valued data.

It is worth noting that these results require no assumptions regarding $g(u, z, \theta)$ (other than measurability). Hence, it transparently covers nonsmooth cases, such as the important case of quantile restrictions. Also, no rank conditions are needed, as we allow for set-identified models.

Remark 2.3 (regarding the choice of ρ) *It is important to realize that the constraints in Definition 2.2 are imposed on the least-favorable family used, not on the true data generating process. As a result, even though each distribution in the selected exponential family admits a moment generating function (in terms of $g(U, z, \theta)$ for a fixed z), this family is able to reproduce the expectation of $g(U, Z, \theta)$ for all distributions of U given Z, including those which do not admit a moment generating function.*

While the method requires a user-specified measure ρ as an input, its choice has absolutely no effect on the results, as long as it satisfies the two conditions stated: (i) its support must match the possible support of the unobservables and (ii) some moment generating function-like quantity must exist and be twice differentiable. The presence of a user-specified measure in the expression of the estimator is analogous to the form of exponential families used in pseudolikelihoods (see Definition 1 in Gourieroux, Monfort, and Trognon (1984)). There is an important difference, however: The choice of the family used in pseudolikelihoods may have an impact on efficiency, whereas the choice of ρ has no effect on the statistical properties of our method. This follows from the fact that Theorem 2.1 and its associated Corollary 2.1 hold for any $\pi \in \mathcal{P}_{\mathcal{Z}}$, not just the true distribution of the observables π_0. In particular, if π is the sample distribution, Corollary 2.1 implies that the range of values spanned by $\hat{g}(\theta, \gamma) \equiv \frac{1}{n}\sum_{i=1}^{n} \tilde{g}(Z_i, \theta, \gamma)$ as γ varies does not depend on ρ. Any objective function based on optimizing a function of $\hat{g}(\theta, \gamma)$ with respect to γ would then have the same value for a given θ, regardless of ρ. In other words, the choice of ρ has no effect on the set $\left\{\hat{g}(\theta, \gamma) : \gamma \in \mathbb{R}^{d_g}\right\}$ in any given finite sample, even though it has an effect on $\tilde{g}(z, \theta, \gamma)$ for a given value of γ. Any specific value of $\hat{g}(\theta, \gamma)$ for one choice of ρ will also be reached by $\hat{g}(\theta, \gamma)$ for another choice of ρ, although perhaps for a different

value of γ.

The presence of an infimum in Equation (6) handles the possibility of a solution "at infinity" ($\|\gamma\| \to \infty$). This happens when the distribution of the unobservables needs to be degenerate at the boundary of $\mathcal{U}$ in order to match the moment conditions. In set-identified models, such solutions often correspond to the boundaries of the identified set for θ and cannot be overlooked. In practice, the presence of solutions "at infinity" in Equation (7) makes little difference, because numerical optimization routines solving for γ abort whenever the objective function no longer changes significantly between iterations. If the solution is "at infinity", these routines will stop at some finite value of γ producing a value of $E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]$ that is close to 0 within a specified tolerance. This is no different from what happens at an interior solution (finite γ), where the optimization routines would stop when producing a value of $E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]$ that is also close to 0 within a specified tolerance. Hence, solutions at infinity do not require a separate treatment in practice.

The condition $E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]=0$ is the first order condition of a convex optimization problem in γ (this follows from $\partial E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]/\partial\gamma'$ being positive-definite, as shown Lemma A.1 of Appendix A), thus making it possible to find γ via numerical routines that are guaranteed to converge. This also implies that any positive-definite quadratic form in $E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]$ will reach its unique global minimum for γ such that $E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]=0$ and has no other local minima (this can be shown by writing $\partial\left(g_\gamma' W g_\gamma\right)/\partial\gamma = 2\left(\partial g_\gamma'/\partial\gamma\right) W g_\gamma$ where $g_\gamma \equiv E_\pi\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]$ and where both $\partial g_\gamma'/\partial\gamma$ and W are positive-definite, so $2\left(\partial g_\gamma'/\partial\gamma\right) W g_\gamma = 0$ iff $g_\gamma = 0$).

2.2 Intuition

We now explain intuitively why Theorem 2.1 would hold. To avoid obscuring the main ideas, we present a heuristic motivation for Theorem 2.1 (Appendix A gives a formal proof).

Given a distribution $\pi \in \mathcal{P}_Z$ of the observable Z, and a $\theta \in \Theta_0$, there may be many possible conditional distributions $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ of the unobservables satisfying the moment condition. Since we only need to find one suitable μ, it is useful to rank the possible μ using some convenient objective function and convert an abstract "existence problem" into a more concrete optimization problem. If there exists no $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ satisfying the moment conditions, the optimization problem will find no solution. If there exists a unique $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ satisfying the moment conditions, the maximization will find it. If there exist more than one $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ (or even infinitely

many) satisfying the moment conditions, the maximization problem will find one of them and it does not matter which one.

For a given $\theta \in \Theta$ and a given marginal distribution of the observables π, the set of all conditional distributions μ (of U given Z) that satisfy the moment conditions is

$$\mathcal{M}_{\theta,\pi} = \left\{ \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}} : E_{\mu\times\pi}\left[g\left(U, Z, \theta\right)\right] = 0 \right\}.$$

To rank distributions in $\mathcal{M}_{\theta,\pi}$ we use entropy, since it has a long history as way to maximize the lack of information under given constraints (Kullback (1959), Shore and Johnson (1980), Csiszar (1991), Golan, Judge, and Miller (1996), Zellner (1997), Imbens, Spady, and Johnson (1998)). This choice may seem arbitrary, but, as we will soon show, it will lead to some remarkable simplifications not possible with other intuitive choices.

Generally, the entropy S of a distribution is defined relative to a reference measure, say, $\rho \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ (which may depend on θ, although this is suppressed in the notation for simplicity) as[10]

$$S\left(\mu||\rho\right) = \begin{cases} -\int\int \ln\left(\frac{d\mu(u|z)}{d\rho(u|z)}\right) d\mu\left(u|z\right) d\pi\left(z\right) & \text{if } \mu \ll \rho \\ -\infty & \text{otherwise} \end{cases}. \tag{8}$$

Among all $\mu \in \mathcal{M}_{\theta,\pi}$ we select the one maximizing this quantity:[11]

$$\mu^{*}\left(\cdot, \theta, \pi\right) = \arg\max_{\mu\in\mathcal{M}_{\theta,\pi}} S\left(\mu||\rho\right).$$

We can set up a Lagrangian for this optimization problem:

$$\begin{aligned} &\int\int \ln\left(f\left(u|z\right)\right) f\left(u|z\right) d\rho\left(u|z\right) d\pi\left(z\right) - \gamma' \int\int g\left(u, z, \theta\right) f\left(u|z\right) d\rho\left(u|z\right) d\pi\left(z\right) \\ &- \int \phi\left(z\right)\left(\int f\left(u|z\right) d\rho\left(u|z\right) - 1\right) d\pi\left(z\right), \end{aligned}$$

where $f\left(u|z\right) \equiv d\mu\left(u|z\right)/d\rho\left(u|z\right)$ and where $\gamma \in \mathbb{R}^{d_g}$ is the Lagrange multiplier vector for the moment constraints while $\phi : \mathbb{R}^{d_z} \mapsto \mathbb{R}$ is the Lagrange multiplier *function* associated with the infinite-dimensional constraint that μ constitutes a valid conditional measure (i.e. $\int d\mu\left(u|z\right) = 1$ or $\int \left(d\mu\left(u|z\right)/d\rho\left(u|z\right)\right) d\rho\left(u|z\right) = 1$ for π-almost every $z \in \mathcal{Z}$). This infinite-dimensional constraint also ensures that the

[10] The notation $\mu \ll \rho$ means that μ admits a density with respect to ρ, which is denoted by Radon-Nikodym derivative $\frac{d\mu(u|z)}{d\rho(u|z)}$.

[11] By convention, we do not exclude a solution μ such that $S\left(\mu||\rho\right) = -\infty$, corresponding to the cases where we do not have $\mu \ll \rho$.

marginal distribution of the observables under $\mu \times \pi$ is equal to π. The first order condition is that the quantity is stationary under small changes in $f(u|z)$, denoted $\delta f(u|z)$:

$$\int \int (1 + \ln f(u|z) - \gamma' g(u, z, \theta) - \phi(z)) \, \delta f(u|z) \, d\rho(u|z) \, d\pi(z) = 0.$$

As this must hold for any $\delta f(u|z)$, we have $1 + \ln f(u|z) - \gamma' g(u, z, \theta) - \phi(z) = 0$, or

$$f(u|z) = \exp(\phi(z) - 1) \exp(\gamma' g(u, z, \theta)). \tag{9}$$

We can solve for $\phi(z)$ by noting that we must have $\int f(u|z) \, d\rho(u|z) = 1$, implying that

$$\exp(\phi(z) - 1) = \left(\int \exp(\gamma' g(u, z, \theta)) \, d\rho(u|z) \right)^{-1}. \tag{10}$$

Substituting (10) in (9), we obtain:

$$f(u|z) = \frac{\exp(\gamma' g(u, z, \theta))}{\int \exp(\gamma' g(u, z, \theta)) \, d\rho(u|z)}. \tag{11}$$

The Lagrange multiplier γ must be such that $\int \int g(u, z, \theta) \, f(u|z) \, d\rho(u|z) \, d\pi(z) = 0$, i.e.

$$\int \int g(u, z, \theta) \frac{\exp(\gamma' g(u, z, \theta))}{\int \exp(\gamma' g(u, z, \theta)) \, d\rho(u|z)} d\rho(u|z) \, d\pi(z) = 0. \tag{12}$$

We have just obtained the expression for the equivalent moment condition stated in Theorem 2.1.

Remark 2.4 *The above reasoning is heuristic, because it overlooks issues such as the validity of the Lagrangian procedure for uncountable constraints or the possibility of solutions "at infinity". It also does not explicitly address the converse result — if θ is not in the identified set, then (12) cannot be satisfied. The proof in Appendix A avoids these issues by directly proving that the existence of a γ solving (12) is equivalent to the original problem of finding at least one distribution of the unobservables that satisfies the moment conditions.*

This heuristic derivation illustrates how the nonparametric problem of the existence of a distribution of the unobservables that satisfies the moment conditions can be reduced to a parametric problem. Initially, we consider any possible distribution and merely rank all valid distributions according to some objective function (here, the entropy). It turns out that the distributions that maximize entropy under given moment constraints form a parametric family that can be indexed by a finite-dimensional

parameter γ. It is well-known within the theory of convex optimization, that the dual of a constraint optimization problem can have a much smaller dimension than the original problem. Here, there is an additional factor in our favor. The number of constraints is infinite in the original problem, so we would have also expected the dual problem to be infinite-dimensional. However, the special form of the entropy functional is such that we can solve for these infinite-dimensional constraints analytically, thus leaving only a finite-dimensional vector γ to solve for numerically.

Note that it is known that, for a finite number of linear constraints, entropy maximization yields a solution whenever there exists at least one distribution satisfying these constraints (e.g. Csiszar (1975), Section 3). However, here, the requirement that the marginal of the observables match the actual observable distribution represents an infinite-dimensional constraint and the standard treatment does not apply.

Using almost any objective function other than entropy would not have resulted in the function $\phi(z)$ nicely separating out in Equation (9), thus precluding an analytic solution. For instance, the most natural alternative would have been maximizing the likelihood $\int\int \ln\left(\frac{d\mu(u|z)}{d\rho(u|z)}\right) d\rho(u|z)\, d\pi(z)$ (instead of (8)). As shown in Section F of the Supplementary Material, this leads to a dual problem where the function $\phi(z)$ enters nonseparably and cannot be solved for analytically. This requires the solution of a different nonlinear optimization problem at each z. Readers familiar with the Empirical Likelihood (EL) literature may be surprised by this result, since the Lagrange multiplier associated with the total unit probability constraint in EL can be solved for analytically. However, applying the same techniques in the present case would require the moment conditions to be satisfied at *each* z, which is not the case in the present case, where they hold after *averaging* over z.

Through calculations similar to those in Section F of the Supplementary Material, it can be shown that any other objective functions associated with the well-known Cressie-Read family (Cressie and Read (1984)) do not admit analytic solutions for $\phi(z)$, except for the objective function $\int\int \left(\frac{d\mu(u|z)}{d\rho(u|z)}\right)^2 d\rho(u|z)\, d\pi(z)$, traditionally associated with the continuous updating GMM estimator. However, this objective function may result in negative probabilities and therefore leads to inconsistent estimates of the identified set in general.[12]

[12] This can be seen in the following simple example: If U is known to be supported on $[-1,1]$, then the identified set for the mean of U is $[-1,1]$. However, if signed measures (still supported on $[-1,1]$) are allowed, then the "mean" could be any real number.

2.3 Estimation outline

The simplest way to evaluate the integral (7) defining the moment function is to draw random vectors u_j, $j = 1, \ldots, R$ from a distribution proportional to $\exp(\gamma' g(u, z, \theta))\, d\rho(u|z;\theta)$ using, e.g., the Metropolis algorithm and calculate the average

$$\hat{g}(z, \theta, \gamma) = \frac{1}{R} \sum_{j=1}^{R} g(u_j, z, \theta). \tag{13}$$

A nice feature of the Metropolis algorithm is that it automatically takes care of the normalization integral in the denominator of (7). This simulation-based approach essentially amounts to plugging-in our least-favorable entropy maximizing family into the method of simulated moments (MSM) (McFadden (1989), Pakes and Pollard (1989)).

As mentioned in Section 2.1, solving for γ includes considering solutions "at infinity". In the limit as $\|\gamma\| \to \infty$, the conditional distribution of the unobservable is typically degenerate and, thanks to the use of an exponential tilting, minimizing the norm of Equation (13) amounts to minimizing a function using the so-called simulated annealing method (Kirkpatrick, Gelatt, and Vecchi (1983)), which is known to be especially effective at avoiding trapping in local minima.

To facilitate optimization with respect to γ or θ, it is useful to construct an average that is a smooth function of θ and γ by construction (provided g is). To this effect, one can exploit the following equality:

$$\tilde{g}(z, \theta, \gamma) = \frac{\int g(u, z, \theta) \exp(\gamma' g(u, z, \theta) - \gamma_0' g(u, z, \theta_0))\; r(u|z, \theta_0, \gamma_0)\, d\rho(u|z;\theta_0)}{\int \exp(\gamma' g(u, z, \theta) - \gamma_0' g(u, z, \theta_0))\; r(u|z, \theta_0, \gamma_0)\, d\rho(u|z;\theta_0)},$$

where

$$r(u|z, \theta_0, \gamma_0) = \frac{\exp(\gamma_0' g(u, z, \theta_0))}{\int \exp(\gamma_0' g(u, z, \theta_0))\, d\rho(u|z;\theta_0)}.$$

For given values of θ_0 and γ_0, one can then evaluate $\tilde{g}(z, \theta, \gamma)$ for any θ, γ by drawing u_j from a density proportional to $\exp(\gamma_0' g(u, z, \theta_0))\, d\rho(u|z;\theta_0)$ and by calculating the ratio of averages:

$$\hat{g}(z, \theta, \gamma) = \frac{\frac{1}{R}\sum_{j=1}^{R} g(u_j, z, \theta) \exp(\gamma' g(u, z, \theta) - \gamma_0' g(u, z, \theta_0))}{\frac{1}{R}\sum_{j=1}^{R} \exp(\gamma' g(u, z, \theta) - \gamma_0' g(u, z, \theta_0))}.$$

Smoothness in the parameters (at least in an almost-everywhere sense) is also important to establish consistency of simulation-based estimators, as it ensures stochastic equicontinuity. The remaining technical complications in the derivation of asymptotic properties associated with the use of simulations to evaluate integrals have been

studied in detail in earlier work (McFadden (1989), Pakes and Pollard (1989), Hajivassiliou and Ruud (1994), Gourieroux and Monfort (1997), Geweke and Keane (2001)). For conciseness, we do not consider such issues here further.

Averaging over the unobservables then provides us with a conventional moment condition $E\left[\tilde{g}\left(Z,\theta,\gamma\right)\right]=0$ involving only observable variables that is equivalent to the original problem. As a result, solving for the parameter θ of interest and for the nuisance parameter γ can be accomplished through a variety of standard techniques. Conventional GMM estimation is perhaps the simplest approach, preferably using the efficient weighting matrix. One-step alternatives to efficient GMM can also be used, such as Empirical Likelihood (EL) or Exponentially Tilted Empirical Likelihood (ETEL), which are known to yield more efficient estimates with a typically smaller small-sample bias in point-identified settings (Newey and Smith (2004), Schennach (2007b)). EL is also known to exhibit desirable optimal power properties under large deviation criteria in the context of point identified (Kitamura (2001), Kitamura, Santos, and Shaikh (2010)) and in a large class of set-identified models (Canay (2010)). While statistical optimality criteria point towards one-step methods, GMM offers one convenient computational advantage: Its objective function involves some sample averages that are linear in $\tilde{g}\left(z,\theta,\gamma\right)$, which enables a more rapid convergence of the simulation-based algorithm (fewer draws of u_j are needed), because averaging over z reduces the noise in $\hat{g}\left(z,\theta,\gamma\right)$.

The possibility of set-identification (rather than point identification) will require special attention when calculating confidence regions. Section G of the Supplementary Material describes how existing general inference techniques (such as Chernozhukov, Hong, and Tamer (2007)) can be used to construct consistent set-estimates and confidence regions and suggests a simple, but conservative, alternative method based on a χ^2 approximation.

2.4 Connection with moment inequalities

An interesting by-product of Theorem 2.1 is that we can rigorously establish that all moment conditions models with unobservables are formally equivalent to moment inequality problems, with the important caveat that *the number of inequalities can be uncountably infinite*. The Models based on Equation (13) in Section D.1 are specific examples of this. In special cases (i.e., when $\text{Closure}\left\{E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right]:\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}\right\}$ is polygonal), this infinite set of inequalities can be reduced to a finite set of inequalities, but not in general.

Theorem 2.2 *The identified set Θ_0 can be equivalently described by*

$$\{\theta \in \Theta : E_{\pi_0}[t(Z,\theta,\eta)] \geq 0 \text{ for all } \eta \in \delta\mathcal{B}_1\},$$

where $\delta\mathcal{B}_1 = \{\eta \in \mathbb{R}^{d_g} : \|\eta\| = 1\}$ *(the unit ball boundary) and*

$$t(Z,\theta,\eta) \equiv \lim_{r\to\infty} \eta' \tilde{g}(Z,\theta,\eta r) \tag{14}$$

for $\tilde{g}(Z,\theta,\gamma)$ *as in Theorem 2.1. Note that if, for some* η*, the limit in (14) diverges then no constraint is associated with this value of* η*. An alternative expression is*

$$t(Z,\theta,\eta) = \sup_{u\in\mathcal{U}} \eta' g(u,Z,\theta). \tag{15}$$

We have already shown (through example 1.3) that the class of models considered here includes models defined via a finite number of moment inequalities as a special case. We now see that it is, in fact, strictly more general than that. While there is some work on infinite sets of moment inequality restrictions[13] (Andrews and Shi (2008), Kim (2008), Menzel (2008), Molinari (2008), Chernozhukov, Lee, and Rosen (2009)), there appears to be little benefit to phrase our class of models entirely in terms of an infinite number of moment inequalities, since our method enables a treatment with a finite-dimensional nuisance parameter and finitely many moment conditions. This connection to moment inequalities also shows, via Equation (15), that our identified set must match the sharp set obtained via inequalities generated from support function methods (Beresteanu, Molchanov, and Molinari (2011) and Ekeland, Galichon, and Henry (2010)) when they apply, while avoiding the often difficult calculation of the support function (Equation (15)), as discussed in Section 3.

3 Relationships to other works

This work touches a number of fields: methods dealing with the presence of unobservables, frameworks to handle set identification, moment inequality models, support function-based convex optimization methods, and information-theoretic methods based on entropy maximization.

A common approach to handle unobservables is the use of a parametric likelihood in which the unobservables are eliminated by integration so that only the marginal distribution of the observables remains. This is conceptually straightforward but crucially

[13] Infinite sets of unconditional moment inequality restrictions ($E[g(Z,\theta,t)] \geq 0 \ \forall t \in \mathcal{T}$) can be cast into conditional moment inequality restrictions ($E[g(Z,\theta,T)|T=t] \geq 0 \ \forall t \in \mathcal{T}$ where T is a random variable uniformly distributed on $\mathcal{T}$).

relies on the ability to correctly specify a fully parametric likelihood, an assumption we wish to avoid.

The method of simulated moments (MSM) (McFadden (1989), Pakes and Pollard (1989)) also performs inference on the basis of a given vector of moment conditions involving unobservables. The MSM proceeds by generating random draws from the distribution of the unobserved variables, assumed to belong to a known parametric family. These draws of the unobservables are combined with the observed data and fed into a conventional Generalized method of Moment (GMM) estimator. This method still requires specifying the distribution of the unobservables, up to a vector of parameters. Our approach is similar in spirit to the MSM but represents the distribution of the unobservables by a carefully constructed least-favorable *parametric* family which is shown to span the exact same range of values of the moment conditions as the corresponding fully nonparametric family. Our method shares the simplicity of the MSM, but entirely eliminates its parametric limitations. The computational requirements of our method are therefore similar to the ones of parametric MSM.

It may be possible to relax the parametric assumptions of the MSM by representing the distribution of the unobservables nonparametrically using a series approximation (see Newey (2001) for an example of this approach). A general asymptotic theory covering this setup in point-identified settings can be found in (Shen (1997)). The difficulty associated with using this approach is the need to let the number of parameters describing the flexible form of the distribution of the unobservables grow with sample size. In contrast, our proposed approach eliminates all parametric distributional assumptions *without* introducing any nuisance parameters whose dimension must increase with sample size, thus providing *significant* computational advantages. For set-identified models, methods based on series approximation would additionally face the problem that the distribution of the unobservables associated with the boundary of the identified set typically exhibit point masses that are difficult to approximate by truncated series of smooth functions.

Our work also has some connections with some previously proposed information-theoretic methods (Shen, Shi, and Wong (1999)) and entropy maximization methods (Golan, Judge, and Miller (1996)), as discussed in more detail in Section J of the Supplementary Material.

We can also make an interesting connection between our approach and models defined via moment inequalities, which have been extensively studied (Chernozhukov, Hong, and Tamer (2007), Andrews and Jia (2008), Chiburis (2008), Rosen (2008), Andrews and Guggenberger (2009), Andrews and Soares (2010), Bugni (2010), Canay

(2010), Romano and Shaikh (2010), Beresteanu, Molchanov, and Molinari (2011)). Many moment condition models involving unobservables are known to imply moment inequality constraints that can be derived by exploiting linearity or monotonicity (see, among many others, Manski (1995), Manski (2003), Magnac and Maurin (2008), Molinari (2008), Example 1.5 above and the general approach of Bontemps, Magnac, and Maurin (2007)). More generally, if, for a given model, the inequalities can be easily derived and their number is finite (or, at the very least, countable), the problem of constructing a suitable objective function has been addressed (notably, in Andrews and Jia (2008), Canay (2010)). However, in general, obtaining equivalent inequalities is not a trivial problem. Our explicit expression (Theorem 2.2) for a set of moment inequalities that is formally equivalent to Equation (1) reveals one important feature: The resulting set of inequalities may be uncountably infinite (even if $g(u, z, \theta)$ and u are both finite-dimensional), thus making methods developed for a finite number of inequalities inapplicable. In constrast, our approach (based on Theorem 2.1) remains finite-dimensional even in moment condition models where the corresponding moment inequality formulation would involve an infinite number of inequalities. Furthermore, we consider not only moment conditions that are linear or monotone in the unobservable, but also arbitrarily complex nonlinear, nonseparable, moment conditions. Analytic tractability of the problem becomes irrelevant when it can be replaced by a generic simulation-based method.

An objective function for moment condition models with unobservables has been suggested in Galichon and Henry (2006) and Ekeland, Galichon, and Henry (2010). Like the present approach, their method manages to replace an infinite-dimensional nuisance parameter by a finite-dimensional one.[14] However, their approach involves the optimization of a nonsmooth function over a bounded set. This entails a number of complications, such as checking for boundary solutions. Their approach amounts to finding a convex hull of what is an intricately "folded" curve or hypersurface in a high-dimensional space in most of the examples we provide in the present paper. As such, their method can be seen as a support function-based methods (Beresteanu, Molchanov, and Molinari (2011)), which can be applied to check if the origin is contained in the convex set of possible values of the moments defining the model. These methods approach the solution along the boundary of the convex set (see right half of Figure 1): At each step, one needs to find the so-called support function, that

[14] A referee pointed out that, even though Equation (3) in Ekeland, Galichon, and Henry (2010) displays a moment that only depends on the unobservables U, their method can cover moment functions that couple the unobservables U and the observables Z, by redefining the unobservables as $U_g \equiv g(U, Z, \theta)$ and using the θ-dependent correspondence $G(u_g, \theta) = \{z : u_g = g(u, z, \theta), u \in \mathcal{U}\}$.

is, the linear inequality that is the closest to the set along a given direction. As this step may involve local extrema issues and boundary solutions, this approach has so far been used only for problems where this step turns out to be simple. This optimization problem is then nested into an outer optimization problem to find the direction of the tightest inequality and check if it is satisfied. While this outer optimization problem has some nice properties (for instance, it is a convex optimization), it is still nonsmooth in general, because there may be kinks at the boundary. Given these difficulties, it is not surprizing that Galichon and Henry (2006) only provide very simple simulation examples of latent variable models where the inequalities are not known in advance (with at most 2 moment conditions and discrete observables). Similarly, Beresteanu, Molchanov, and Molinari (2011) focuses on examples where the support function is a maximum over a discrete set of at most 3 elements and where the observable variables are discrete. In both cases, the (typically difficult) inner optimization problem (over the unobservables) is simple, and only the outer optimization to find the tightest inequality remains and has to be performed a small number of times.

In contrast (see left half of Figure 1), the current approach instead results in an objective function that is smooth in the finite-dimensional nuisance parameter γ. This simplification is made possible through the following realization. Instead of devoting considerable effort in obtaining numerically exact inequalities defining the convex hull of possible values of the moments (as defined in Corollary 2.2) and then checking whether it contains the origin, the proposed method parametrizes the interior of this convex hull via a smooth function (of γ) that can be inexpensively calculated from simple moments, thus enabling the verification of whether the origin is included in the convex hull via standard smooth optimization methods that approach the solution from the "inside" of the convex hull rather than along its potentially nonsmooth boundary.[15] Thanks to these simplifications, our examples in Section 5 include up to 27 moment conditions, with all observed and unobserved variables being continuous.

Another limitation of Galichon and Henry (2006) and Ekeland, Galichon, and Henry (2010) is that their method does not cover unbounded moment functions[16]

[15] When the convex hull does not contain the origin, our method amounts to finding the point inside the convex hull that is the closest to the origin via an optimization method known as simulated annealing (Kirkpatrick, Gelatt, and Vecchi (1983)), which is known to be especially effective at avoiding trapping in local minima. Our approach also bears some resemblance with so-called "interior point methods" in convex optimization (Boyd and Vandenberghe (2004), Chap. 11).

[16] One of their optimization steps requires compactness of the range of the moment functions to ensure that the minimum is not at $-\infty$ for nonzero values of the Lagrange multiplier of the moment constraints, which could mask the true optimum.

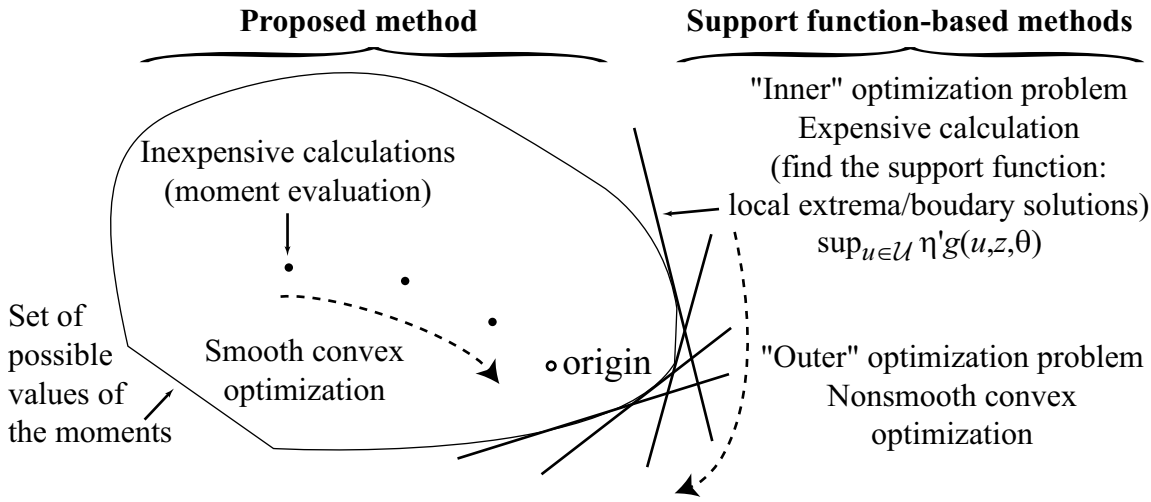

Figure 1: Comparison of the proposed method with methods based on support functions (Galichon and Henry (2006), Ekeland, Galichon, and Henry (2010), Beresteanu, Molchanov, and Molinari (2011)). (To reduce the dimensionality of the problem so that it can be pictured, this figure considers a simple case where the observable Z is constant, and where the number of moment conditions is only 2.)

(such as the mean of a variable with unbounded support, a fairly common occurence). Similarly, Beresteanu, Molchanov, and Molinari (2011) also work with sets that are bounded (with probability one). Among the examples provided in the present paper, only the interval-valued data example solely involves bounded sets. Finally, in specific cases where a natural choice of objective function already exists (such as moment inequality models or conventional underidentified GMM), the value of Galichon and Henry's objective function outside of the identified set does not necessarily coincide with any of the existing results. While this is not needed for their method to be valid, it would be conceptually desirable. In contrast, our approach has the property that it can nest the objective function of GMM or any of its one-step alternatives (GEL, ETEL) as special cases.

Although the ELVIS approach is very general, it does not mean it should always be the preferred method. Naturally, if some of the steps leading to the identified set can be carried out analytically at a modest effort, then this would likely lead to lower computational requirements (this may happen, for instance, if the support function can easily be computed analytically and if the resulting optimization over inequalities is well-behaved).

4 Generalizations

4.1 Flexible supports

The set $\mathcal{U}$ may, in general, depend on z or θ, but such a dependence can always be eliminated by rewriting an equivalent estimation problem in which the dependence of the support $\mathcal{U}$ on z or θ has been incorporated into the function $g(u, z, \theta)$. Specifically, let $g(u, z, \theta)$ denote the original moment conditions and let $\mathcal{U}_{z,\theta}$ denote the z- and θ-dependent support of U. Consider a set $\overline{\mathcal{U}}$ having a cardinality larger or equal to the cardinality of any of $\mathcal{U}_{z,\theta}$. Construct a many-to-one (which may be reduced to one-to-one) and onto (z, θ)-dependent measurable mapping $m(\cdot, z, \theta) : \overline{\mathcal{U}} \mapsto \mathcal{U}_{z,\theta}$. Define a new moment condition as $E\left[\mathring{g}\left(\mathring{U}, Z, \theta\right)\right] = 0$ where $\mathring{g}(\mathring{u}, z, \theta) = g(m(\mathring{u}, z, \theta), z, \theta)$ and where $\mathring{U}$ is a random variable having support $\mathcal{U} \equiv \overline{\mathcal{U}}$ that does not depend on z or θ. Hence, the z- and θ-dependent support case can be reduced to the constant support case. It follows that all of our results trivially continue to hold with $\mathcal{U}$ replaced by $\mathcal{U}_{z,\theta}$. A constant set $\mathcal{U}$ simplifies the exposition and results in no loss of generality. It avoids replacing simple quantities such as $\mathcal{U} \times \mathcal{Z}$ and $\mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ by much less transparent counterparts. In the implementation of the method, it may be more convenient to use $\mathcal{U}_{z,\theta}$ and keep the original function $g(u, z, \theta)$.

4.2 Nonlinear functions of moments

Some applications necessitate moment constraints that involve nonlinear functions of expectations. For instance, independence between two random quantities $g_1(U, Z, \theta)$ and $g_2(U, Z, \theta)$ implies the moment condition:

$$E[g_1(U, Z, \theta) g_2(U, Z, \theta)] - E[g_1(U, Z, \theta)] E[g_2(U, Z, \theta)] = 0.$$

This constraint can be readily converted into a set of constraints that are linear in the expectations by introducing a nuisance parameter ϕ:

$$\begin{aligned} E[g_1(U, Z, \theta) - \phi] &= 0 \\ E[g_1(U, Z, \theta) g_2(U, Z, \theta) - \phi g_2(U, Z, \theta)] &= 0. \end{aligned}$$

This approach can be fully generalized. If

$$f(E[g(U, Z, \theta)]) = 0$$

for some nonlinear function $f : \mathbb{R}^{d_g} \mapsto \mathbb{R}^{d_f}$, then one can introduce a nuisance parameter vector $\phi \in \mathbb{R}^{d_g}$ and equivalently write linear moment conditions:

$$E[g(U, Z, \theta) - \phi] = 0$$

with the expanded parameter space being $\Theta^* = \Theta \times \Phi$ where

$$\Phi = \left\{\phi \in \mathbb{R}^{d_g} : f(\phi) = 0\right\}.$$

Inference regarding θ can then be carried out using an objective function where the nuisance parameters ϕ have been "profiled" out (see Section G.1 in the Supplementary Material).

4.3 Conditional moments, independence restrictions and state-space models

It is natural to consider the extension of conventional moment restrictions to conditional moment restrictions of the form

$$E\left[g(U, Z, \theta) \left| c(U, Z, \theta)\right.\right] = 0 \tag{16}$$

with probability 1, for two given functions g and c. It is well known that conditional moment restrictions of the form (16) are equivalent to an infinite family of unconditional moment restrictions (Chamberlain (1987), Bierens (1990), Stinchcombe and White (1998))

$$E\left[g(U, Z, \theta)\, s\left(c(U, Z, \theta), t\right)\right] = 0,$$

where $s(\cdot, t)$ is a suitable family of functions indexed by t. The index t can be discrete, because a countable set of unconditional moments is sufficient to impose a conditional mean restriction (see, for instance, Chamberlain (1987), pp. 324–325, or Proposition C.1 in Section C of the Supplementary Material).

A similar idea can be used to enforce independence restrictions. If one wishes to specify that $a(U, Z, \theta)$ is independent from $b(U, Z, \theta)$, one could impose an infinite set of moment factorization constraints (indexed by t and $\tilde{t}$):

$$E\left[s\left(a(U, Z, \theta), t\right) s\left(b(U, Z, \theta), \tilde{t}\right)\right] = E\left[s\left(a(U, Z, \theta), t\right)\right] E\left[s\left(b(U, Z, \theta), \tilde{t}\right)\right], \tag{17}$$

where $s(\cdot, t)$ is a suitable family of functions. Such nonlinear functions of moments can be converted into an equivalent sequence of moment conditions $E\left[g_j(u, z, \theta, \nu)\right] = 0$ for $j = 1, 2, \ldots$ via the introduction of nuisance parameters ν using the techniques of Section 4.2. The equivalence between independence and a sequence of moment factorization constraints is shown formally in Proposition C.2 of Section C of the Supplementary Material.

We now state an infinite dimensional version of Theorem 2.1 which covers all of the above situations.

Definition 4.1 *For all $\theta \in \Theta$, any $\nu \in \mathcal{V}$ and any $J \in \mathbb{N}^*$, let $\rho\left(\cdot|\cdot;\theta,\nu^{(J)}\right) \in P_{\mathcal{U}|\mathcal{Z}}$ be a user-specified conditional measure satisfying (i) $\operatorname{supp}\rho\left(\cdot|z;\theta,\nu^{(J)}\right) = \mathcal{U}$ for each $z \in \mathcal{Z}$ and (ii) $E_\pi\left[\ln E_{\rho\left(\cdot|\cdot;\theta,\nu^{(J)}\right)}\left[\exp\left(\sum_{j=1}^J \gamma_j g_j(u,z,\theta,\nu)\right)|Z\right]\right]$ exists and is twice differentiable in γ for all $\gamma \in \mathbb{R}^J$.*

Theorem 4.1 *Let $E[g_j(u,z,\theta,\nu)] = 0$ for $j = 1,2,\ldots$ be a sequence of moment restrictions potentially depending on a vector of nuisance parameter $\nu \in \mathcal{V}$, where the set $\mathcal{V}$ may be infinite-dimensional, but $g_j(u,z,\theta,\nu)$ only depends on a finite number of elements of ν. Let*

$$\Theta_0 = \left\{\theta \in \Theta : \inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \sup_{j\in\mathbb{N}^*} |E_{\mu\times\pi_0}[g_j(U,Z,\theta,\nu)]| = 0\right\}$$

and

$$\Theta_0^{(J)} = \left\{\theta \in \Theta : \inf_{\nu^{(J)}\in\mathcal{V}^{(J)}} \inf_{\gamma\in\mathbb{R}^J} \left\|E_\pi\left[\tilde{g}^{(J)}\left(Z,\theta,\nu^{(J)},\gamma\right)\right]\right\| = 0\right\},$$

where

$$\tilde{g}^{(J)}\left(Z,\theta,\nu^{(J)},\gamma\right) \equiv \frac{\int g^{(J)}\left(u,z,\theta,\nu^{(J)}\right)\exp\left(\gamma' g^{(J)}\left(u,z,\theta,\nu^{(J)}\right)\right)d\rho\left(u|z;\theta,\nu^{(J)}\right)}{\int \exp\left(\gamma' g^{(J)}\left(u,z,\theta,\nu^{(J)}\right)\right)d\rho\left(u|z;\theta,\nu^{(J)}\right)},$$

where $\rho\left(\cdot|\cdot;\theta,\nu^{(J)}\right)$ is as in Definition 4.1 and $g^{(J)}\left(u,z,\theta,\nu^{(J)}\right) = \left[g_j\left(u,z,\theta,\left(\nu^{(J)},0\right)\right)\right]_{j=1}^J$ in which $\nu^{(J)} \in \mathcal{V}^{(J)}$ denotes the elements on $\nu \in \mathcal{V}$, upon which $g^{(J)}\left(u,z,\theta,\nu^{(J)}\right)$ depends. Then, (i) $\Theta_0^{(J+1)} \subseteq \Theta_0^{(J)}$, (ii) $\cap_{J\in\mathbb{N}^*}\Theta_0^{(J)} = \Theta_0$ and (iii) $d_H\left(\Theta_0^{(J)},\Theta_0\right) \to 0$ as $J \to \infty$, where $d_H(\mathcal{A},\mathcal{B}) \equiv \max\left\{\sup_{\alpha\in\mathcal{A}}\inf_{\beta\in\mathcal{B}}\|\alpha-\beta\|, \sup_{\beta\in\mathcal{B}}\inf_{\alpha\in\mathcal{A}}\|\alpha-\beta\|\right\}$ is the Hausdorff metric:

Typically, this method would be implemented by letting J grow with sample size, as is commonly done in conditional moment models (e.g. Donald, Imbens, and Newey (2008), in the case of fully observed variables). Although the above identification result holds for any rate of divergence of J to infinity, performing inference may require a controlled growth rate for J (to maintain the π_0-Donsker property of sample averages of the moment functions). The well-known semiparametric efficiency result of Chamberlain (1987) (i.e. there exists a finite vector of unconditional moment constraints yielding the same efficiency as the original conditional moments), suggests that, in finite samples, the loss of efficiency associated with replacing an infinite number of constraints by a finite number of moment constraints may be small.

Allowing for an infinite dimensional nuisance parameter ν is essential to cover independence constraints. It is not needed for conditional moment restrictions (in

which case Theorem 4.1 applies with $\mathcal{V}$ reduced to a singleton). Theorem 4.1 covers not only conditional mean and independence, but also any other constraints that can be phrased as a sequence of moment conditions. Another example of such infinite-dimensional restrictions is the equality between the marginal distributions of different unobservables. This type of restriction could be useful, for instance, in dynamic state-space models (e.g. Harvey, Koopman, and Shephard (2004), Harvey (2004)), where distributional assumptions could be replaced by moment conditions that may involve coupling between two different lags v_{t-l} and v_t of the same stationary sequence of some unobservable variable v_t. In such cases, one may need to introduce a two-dimensional unobservable, i.e. $u = (v_{t-l}, v_t)$ to impose a constraint of the form $E[v_{t-l}v_t] = 0$, where it must be ensured that v_{t-l} and v_t have the same marginal distribution, if the process is stationary.

The general class of models covered by Theorem 4.1 also admits, as special cases, all models defined via a countable number of moment inequalities, through the device introduced in Example 1.3. One complication associated with these generalizations is that the treatment does involve an optimization problem whose dimensionality grows with sample size, unlike the simpler case of unconditional moment constraints. Nevertheless, the dimensionality of the quantities involved is smaller than other existing methods that could plausibly be adapted to this setting. Series approximations to the unobservable distribution would generally require the number of nuisance parameter per moment condition to go to infinity as sample size grows (while this ratio remains finite with our method). Similarly, moment inequality methods (constructed, e.g., via Theorem 2.2) would require an infinite number of inequalities even for finite J.

Remark 4.1 *A nice feature of Theorem 4.1 is that the approximate identified set $\Theta^{(J)}$ obtained with a finite number of constraints is slightly conservative. Therefore, inference based on this approach would be strictly valid (although conservative) in finite samples, which is considerably better than a method that would reject the null too often in finite sample, thus giving an illusion of accuracy. The latter situation would arise, for instance, if one were to merely write a likelihood function for the model in terms of nonparametric unobservable densities approximated by truncated series. In a finite sample, the parametric assumptions involuntarily implied by truncation of the series would tend to bias the size of the identified set systematically downward. In contrast, our approach always includes least-favorable distributions by construction and provides reliable conservative confidence regions that approach the true identified set "from the outside" (rather than "from the inside").*

Remark 4.2 *While it is conceptually straightfoward to formally establish the validity of resampling/subsampling methods in the case where the number of moment condition is finite (e.g., using the methods in Chernozhukov, Hong, and Tamer (2007), as explained in Section G of the Supplementary Material), it is technically nontrivial to do so when the number of moment condition increases with sample size (as it does for the extension considered here). However, such difficulties are not specific to ELVIS and, in fact, often occur in nonparametric or semiparametric asymptotic analysis.*

5 Simulations

5.1 Interval-valued data and censored regression

Sections D.1 and D.2 of the Supplementary Material describe in detail simulation experiments based on our Examples 1.1 (regression with interval-valued data) and 1.2 (censored regression), respectively. They illustrate some key features of the method. First, the set over which the objective function vanishes matches the well-known bounds for these models (this is also verified analytically in Section I of the Supplementary Material for Example 1.1). Second, one can easily add plausible moment conditions to narrow down the identified set. In these examples, the worst-case scenario giving rise to the bounds may be associated with implausible patterns of heteroskedasticity in the residuals that can be restricted by adding moment conditions ensuring that the variance of the residuals is not correlated with the regressors or their magnitude. As shown in Figure 2, the reduction in the identified set is particularly striking in the censored example. Interestingly, handling these more complex models requires no additional effort on the part of the researcher (even though the moment functions are nonmonotone in the unobservables, which would make an analytic solution difficult) — the simulations take care of everything.

5.2 Errors-in-variables models

We first consider the simplest errors-in-variables model of Example 1.5, with a sample of 250 iid observations generated with $X^* \sim N(0,1)$, $V_1 \sim N(0,1/4)$ and $V_2 \sim N(0,1/4)$. The algorithm of Section 2.3 (with empirical likelihood) was used with $R = 2000$, after 100 equilibration steps. In Figure 3, the objective function is seen to agree very well with the usual standard "forward and reverse regression" bounds

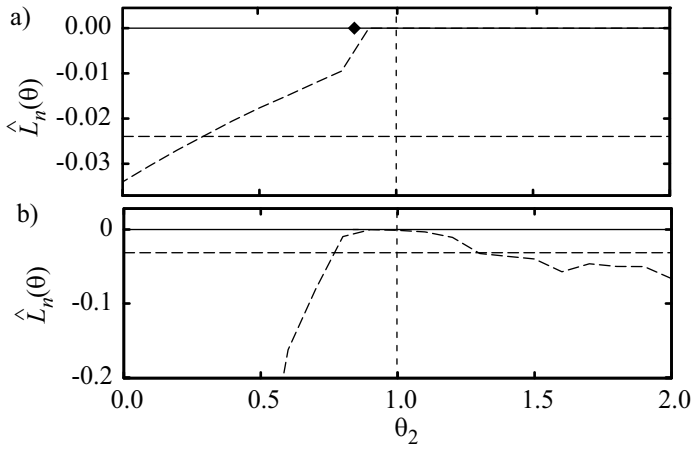

Figure 2: Objective function for a censored regression model. a) Result obtained with the usual uncorrelatedness and zero mean assumptions on the residuals. The upper diamond mark the well-known analytic lower bound for this model. b) Same exercise while assuming, in addition, that the variance of the 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.

(e.g. Klepper and Leamer (1984)) for this model.[17]

We can build upon this simple model. It is known that a linear specification with all variables normally distributed is at best set identified but that point identification is possible when the regressor is not normally distributed and when X^*, V_1 and V_2 are mutually independent. These are ideal test cases because they illustrate the method's ability to transparently cover both set- and point-identified models. We use a sample of 250 iid observations generated as in Example 1.5 with $V_1 \sim N(0, 1/4)$ and $V_2 \sim N(0, 1/4)$ We consider the case (i) where $X^* \sim N(0, 1)$ and (ii) where X^* is drawn from a uniform distribution with zero mean and unit variance.

Example 1.5 (continued) *Mutual independence between X^*, V_1 and V_2 can be imposed via a sequence of moment factorization constraints, as described in Section 4.3. Here, we require V_1 and V_2 to have zero mean and all mixed moments of X^*, V_1 and V_2 up to order 4 to factor, e.g., $E\left[(X^*)^2 (V_1)^2\right] = E\left[(X^*)^2\right] E\left[(V_1)^2\right]$. This involves using the techniques described in Section 4.2 and necessitates the introduction of three nuisance parameters $\theta_2, \theta_3, \theta_4$, so that $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$. Specifically, the vector of*

[17] In fact, the objective function obtained from the sample should be exactly zero between these bounds in this case, but a small residual numerical noise is visible here. While these fluctuations can be virtually eliminated by tightening the optimization tolerance and simulating the unobservables for a longer time, it is unnecessary to do so, because these fluctuations become inconsequential when they are orders of magnitude smaller than the critical value.

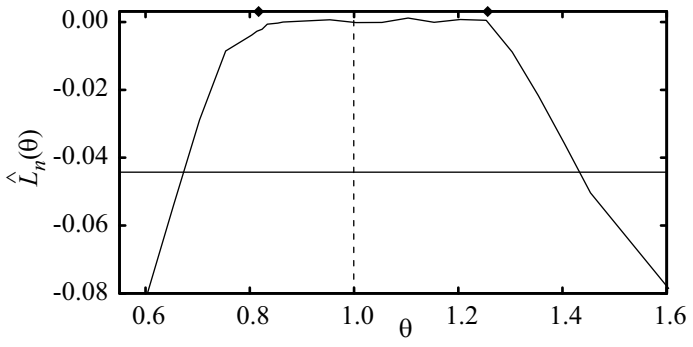

Figure 3: Objective function for a simple measurement error model assuming mutual uncorrelatedness between the true regressor and the two errors. The upper diamonds mark the standard "forward and reverse regression" bounds for this model. The horizontal line indicates the critical value (at the 95% level) and the true value of the parameter is indicated by a vertical dashed line.

moment conditions has 27 elements: $g(U, Z, \theta) = (X^* - \theta_2,\ X^{*2} - \theta_3,\ V_1^2 - \theta_4,\ V_1,\ V_2,\ X^*V_1,\ X^*V_2,\ V_1V_2,\ X^{*2}V_1,\ X^*V_2,\ V_1^2(X^* - \theta_2),\ V_2^2(X^* - \theta_2),\ X^*V_1V_2,\ X^{*3}V_1,\ X^{*3}V_2,\ V_1^3V_2,\ V_2^3V_1,\ V_1^3(X^* - \theta_2),\ V_2^3(X^* - \theta_2),\ (X^{*2} - \theta_3)V_1^2,\ (X^{*2} - \theta_3)V_2^2,\ (V_1^2 - \theta_4)V_2^2,\ X^{*2}V_1V_2,\ V_1^2X^*V_2,\ V_2^2X^*V_1)'$. *The number of unobservables is still 1, because* V_1 *and* V_2 *can be expressed in terms of* X^* *and the observable variables (*$V_1 = X - X^*$ *and* $V_2 = Y - X^*\theta$*).*

While it is known that it is possible to analytically construct a set of moment restrictions that exploit the information provided by moments up to 4 (Cragg (1997)), our method provides an equivalent way to do this while bypassing most of the difficult analytical work. Figure 4 compares the objective functions obtained for the set-identified normal case and the point-identified uniform case. The nuisance parameters $(\theta_2, \theta_3, \theta_4)$ are profiled out. Note that the variance of X^* is the same in both subcases to ensure that the results are indeed driven by information provided by the higher-order moments and not by changes in the second moments of the data. The point-identified case exhibits a clearly more localized maximum in its objective function. In contrast, the objective function in the set-identified case displays a flatter region which is consistent with the usual bounds, indicated by diamonds. In this case, the objective function is not perfectly flat between the bounds and this situation persists as the numerical accuracy of the calculations is improved. The objective function over the identified set is clearly at a finite distance from zero, which is a clear indication that the model is over-identified (all moment conditions need not be satisfied in a given sample). This is not pathological — it is merely a clear indication that the model

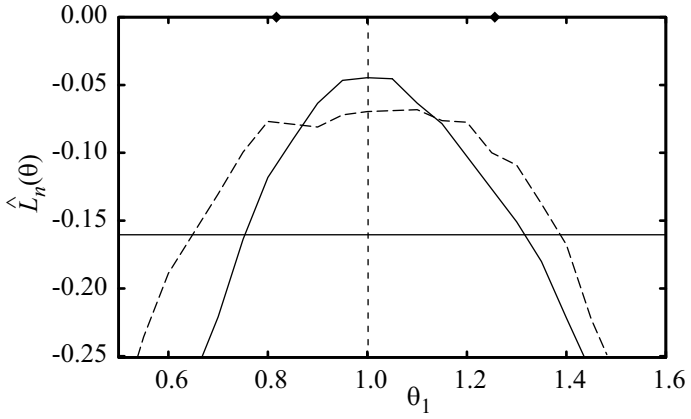

Figure 4: Objective function for the measurement error model under the assumption of mutual independence between the regressors and the two errors. The dashed curve is for the set-identified case with normally distributed regressor while the solid curve is for the point-identified case with uniformly distributed regressor. The horizontal line as the critical value (at the 95 % level), the topmost diamonds mark the usual "forward and reverse regression" bounds for this model and the true value of the parameter is indicated by a vertical dashed line..

happens to not satisfy the so-called degeneracy property (Chernozhukov, Hong, and Tamer (2007)). As such, this model provides an important example of a model that is set-identified and yet, over-identified.[18]

Although the shapes of the objective functions are very revealing regarding the nature of the identification (point- or set-identification), it is interesting to note that the lengths of the corresponding confidence regions are not strikingly different. This reflects the fact that the identification power provided by the higher-order moments in linear specifications can be somewhat "weak", an issue that has been observed in some applications (Hausman, Newey, and Powell (1995)). This problem becomes more severe as the distribution of the regressors approaches normality.

Section D.3 of the Supplementary Material considers a nonlinear errors-in-variable model without side information, that is, Example 1.5 for a nonlinear specification $Y = r(X^*, \theta) + V_2$ where $r(X^*, \theta)$ has one of the two following forms:

$$r(X^*, \theta) = \theta_1 X^* + \theta_2 (X^*)^2 \tag{18}$$

$$r(X^*, \theta) = \theta_1 X^* + \theta_2 \exp(X^*). \tag{19}$$

[18] Other examples are easy to construct: Combine a moment inequality model involving a subset of the parameters with an overidentified moment condition model involving another subset of the parameters. In our example, the under- and over-identified components cannot be so easily separated.

While it has recently been shown that such models can be point-identified under full mutual independence assumptions (Schennach and Hu (2013)), no such result exists under weaker uncorrelatedness conditions. 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 (19). In contrast, our method applies directly — only trivial changes in the program handling the standard measurement error problem were needed.

6 Conclusion

This paper introduces a generalization of GMM to moments involving unobservable variables that circumvents the need for infinite-dimensional nuisance parameters. The key idea is to model the distribution of the unobservables via an entropy-maximizing least-favorable parametric family of distributions that exactly reproduces the same range of moment values as the original nonparametric family. The resulting feasible moment conditions can be used within a GMM framework (or any of its one-step alternatives) and transparently cover both point- and set-identified models. Extensions to conditional moments, independence restrictions and some state-space models are also given.

A Proofs

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).

Lemma A.1 *Let Assumption 2.1 hold and assume that for all $\theta \in \Theta$, any unit vector η and for all z in some subset of $\mathcal{Z}$ of positive probability (under the measure π), $\inf_{u\in\mathcal{U}} \eta' g(u,z,\theta) \neq \sup_{u\in\mathcal{U}} \eta' g(u,z,\theta)$. Then, for ρ as in Definition 2.2,*

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

and

$$V_\gamma = \int \frac{\int (g(u,z,\theta)-\tilde{g}(z,\theta,\gamma))(g(u,z,\theta)-\tilde{g}(z,\theta,\gamma))'\exp(\gamma' g(u,z,\theta))\, d\rho(u|z)}{\int \exp(\gamma' g(u,z,\theta))\, d\rho(u|z)} d\pi(z) \tag{20}$$

are such that, at each $\gamma \in \mathbb{R}^{d_g}$, $\|g_\gamma\| < \infty, \|V_\gamma\| < \infty$ *and* V_γ *is positive-definite. Moreover, derivatives with respect to* γ *up to order* 2 *commute with the expectations in* $E_\pi \left[\ln E_\rho \left[\exp\left(\gamma' g\left(U, Z, \theta\right)\right) | Z\right]\right]$ *and* $V_\gamma = \partial g_\gamma / \partial \gamma'$.

Proof. See Section C of the Supplementary Material. ■

Lemma A.2 *Given a probability measure* F *of a nondegenerate random variable* X *taking values in* $\mathbb{R}$, *if for any* $\lambda \in \mathbb{R}$, $M\left(\lambda\right) \equiv \int x \exp\left(\lambda x\right) dF\left(x\right) / \int \exp\left(\lambda x\right) dF\left(x\right)$ *exists, then for all* $\lambda \in \mathbb{R}$,

$$M\left(\lambda\right) < \lim_{\lambda \to \infty} M\left(\lambda\right) = \sup \operatorname{supp} F,$$

where the right-hand side (the upper bound of the support of F*) could be infinite.*

Proof. Let $\mathcal{H}$ denote the convex hull of the support of F (i.e. the smallest closed interval containing the support of F). Let b be any point in the interior of $\mathcal{H}$ (which is nonempty by assumption). Without loss of generality, we may assume that $b > 0$ (Otherwise just add the same constant to x, X and b and note that $M\left(\lambda\right)$ and $\sup \operatorname{supp} F$ would just be shifted by that same constant, since the multiplicative shift in the exponentials cancels in the ratio defining $M\left(\lambda\right)$). The conclusion is equivalent to showing that, for λ sufficiently large, $M\left(\lambda\right)$ will eventually exceed b. Let $c = b + \varepsilon$, with $\varepsilon > 0$ small enough so that $b + \varepsilon$ is still inside of $\mathcal{H}$. We then have

$$\begin{aligned}
M\left(\lambda\right) &= \frac{\int_{x<c} x \exp\left(\lambda x\right) dF\left(x\right) + \int_{x \geq c} x \exp\left(\lambda x\right) dF\left(x\right)}{\int \exp\left(\lambda x\right) dF\left(x\right)} \\
&\geq \frac{-\int_{x<c} |x| \exp\left(\lambda x\right) dF\left(x\right) + \int_{x \geq c} x \exp\left(\lambda x\right) dF\left(x\right)}{\int \exp\left(\lambda x\right) dF\left(x\right)} \\
&\geq \frac{-\int_{x<c} |x| \, dF\left(x\right) \exp\left(\lambda c\right) + c \int_{x \geq c} \exp\left(\lambda x\right) dF\left(x\right)}{\int \exp\left(\lambda x\right) dF\left(x\right)} \\
&= \frac{-\int_{x<c} |x| \, dF\left(x\right) \exp\left(\lambda c\right) + c \int_{x \geq c} \exp\left(\lambda x\right) dF\left(x\right)}{\int_{x<c} \exp\left(\lambda x\right) dF\left(x\right) + \int_{x \geq c} \exp\left(\lambda x\right) dF\left(x\right)} \\
&= \frac{-\int_{x<c} |x| \, dF\left(x\right) + c \left[\int_{x \geq c} \exp\left(\lambda\left(x - c\right)\right) dF\left(x\right)\right]}{\int_{x<c} \exp\left(\lambda\left(x - c\right)\right) dF\left(x\right) + \left[\int_{x \geq c} \exp\left(\lambda\left(x - c\right)\right) dF\left(x\right)\right]}. \qquad (21)
\end{aligned}$$

Note that the terms in bracket can be shown to diverge: For some $\varepsilon > 0$ such that $c + \varepsilon$ is still inside $\mathcal{H}$,

$$\int_{x \geq c} \exp(\lambda(x-c))\, dF(x) \geq \int_{x \geq c+\varepsilon} \exp(\lambda(x-c))\, dF(x)$$
$$\geq \exp(\lambda(c+\varepsilon-c)) \int_{x \geq c+\varepsilon} dF(x) = \exp(\lambda\varepsilon) \int_{x \geq c+\varepsilon} dF(x) \to \infty$$

as $\lambda \to \infty$ since $\int_{x \geq c+\varepsilon} dF(x) > 0$ as $c+\varepsilon$ is in $\mathcal{H}$. It follows that for sufficiently large λ, the numerator of (21) is positive and we can write (because $\int_{x<c} \exp(\lambda(x-c))\, dF(x) \leq \int_{x<c} dF(x) \leq 1$)

$$\frac{\int x \exp(\lambda x)\, dF(x)}{\int \exp(\lambda x)\, dF(x)} \geq \frac{-\int_{x<c} |x|\, dF(x) + c\int_{x \geq c} \exp(\lambda(x-c))\, dF(x)}{1 + \int_{x \geq c} \exp(\lambda(x-c))\, dF(x)} \to c > b.$$

Hence, we have shown any b in the interior of $\mathcal{H}$ will eventually be exceed as $\lambda \to \infty$. This, combined with the fact that $M(\lambda)$ can only yield a value inside $\mathcal{H}$ for finite λ, concludes the proof. ■

Proof of Theorem 2.1. For given $\theta \in \Theta$ and $\pi \in \mathcal{P}_{\mathcal{Z}}$, let

$$\mathcal{K}_{\theta,\pi} = \text{Closure}\left\{E_{\mu\times\pi}[g(U,Z,\theta)] : \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}\right\},$$

the closure of the set of all possible moment values. Note that $\mathcal{K}_{\theta,\pi}$ is convex because if $\kappa_{1,j} \equiv E_{\mu_{1,j}\times\pi_0}[g(U,Z,\theta)]$ and $\kappa_{2,j} \equiv E_{\mu_{2,j}\times\pi_0}[g(U,Z,\theta)]$ are both sequences converging in $\mathcal{K}_{\theta,\pi}$, then $\kappa_{3,j} = \omega\kappa_{1,j} + (1-\omega)\kappa_{2,j}$ for any $\omega \in [0,1]$ also does because it can be generated from the sequence of measures $\mu_{3,j} = \omega\mu_{1,j} + (1-\omega)\mu_{2,j} \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ through $\kappa_{3,j} = E_{\mu_{3,j}\times\pi_0}[g(U,Z,\theta)]$.

Without loss of generality, we assume that, for all $\theta \in \Theta$, any unit vector η and for all z in some subset $\mathcal{Z}_+$ of $\mathcal{Z}$ that has positive probability under the measure π, $\inf_{u\in\mathcal{U}} \eta' g(u,z,\theta) \neq \sup_{u\in\mathcal{U}} \eta' g(u,z,\theta)$. If that is not the case for some η, this means some linear combination of moment conditions does not depend on u. A suitable linear transformation of $g(u,z,\theta)$, would then produce an equivalent moment vector of the form:

$$\begin{bmatrix} g_u(u,z,\theta) \\ g_z(z,\theta) \end{bmatrix}.$$

Equation (7) then becomes:

$$\tilde{g}\left(z,\theta,(\gamma_u',\gamma_z')'\right) \equiv \begin{bmatrix} \frac{\int g_u(u,z,\theta)\exp(\gamma_u' g_u(u,z,\theta)+\gamma_z' g_z(z,\theta))d\rho(u|z)}{\int \exp(\gamma_u' g_u(u,z,\theta)+\gamma_z' g_z(z,\theta))d\rho(u|z)} \\ \frac{\int g_z(z,\theta)\exp(\gamma_u' g_u(u,z,\theta)+\gamma_z' g_z(z,\theta))d\rho(u|z)}{\int \exp(\gamma_u' g_u(u,z,\theta)+\gamma_z' g_z(z,\theta))d\rho(u|z)} \end{bmatrix} = \begin{bmatrix} \frac{\int g_u(u,z,\theta)\exp(\gamma_u' g_u(u,z,\theta))d\rho(u|z)}{\int \exp(\gamma_u' g_u(u,z,\theta))d\rho(u|z)} \\ g_z(z,\theta) \end{bmatrix}.$$

The dependence on γ_z disappears and the subvector $g_z(z,\theta)$ is unaffected by the averaging, that is, it behaves like a regular moment condition that does not require any special treatment to establish its identifying power. Hence, we focus our attention on $g_u(u,z,\theta)$, which we rename $g(u,z,\theta)$, and assume that all moment conditions do depend on u, that is, $\inf_{u\in\mathcal{U}} \eta' g(u,z,\theta) \neq \sup_{u\in\mathcal{U}} \eta' g(u,z,\theta)$ for all $\theta \in \Theta$, any unit vector η and for all $z \in \mathcal{Z}_+$.

Let $g_\gamma \equiv E_\pi[\tilde{g}(Z,\theta,\gamma)]$ and $V_\gamma \equiv \frac{\partial g_\gamma}{\partial \gamma'}$ be defined as in Lemma A.1. For a given $\theta \in \Theta$ and a given $\pi \in \mathcal{P}_\mathcal{Z}$, we wish to show that $0 \in \mathcal{K}_{\theta,\pi}$ iff there exists a path $\gamma : \mathbb{R}^+ \mapsto \mathbb{R}^{d_g}$ such that $\lim_{t\to\infty} g_{\gamma(t)} = 0$. Note that we allow for solutions "at infinity" (i.e. $\|\gamma(t)\| \to \infty$ as $t \to \infty$).

We start with a trial value of $\gamma = 0$ and gradually update γ via a differential equation until the moment conditions are satisfied. Specifically, set $\gamma(0) = 0$, and update $\gamma(t)$ as the parameter t increases according to

$$\frac{d\gamma(t)}{dt} = -\frac{1}{2} V_{\gamma(t)}^{-1} g_{\gamma(t)}. \tag{22}$$

By Lemma A.1, the interchanges of differentiation and integration performed above are justified and g_γ, V_γ and V_γ^{-1} exist for all finite values of γ, then

$$\begin{aligned} \frac{d}{dt}\left\|g_{\gamma(t)}\right\|^2 &= \frac{d}{dt}\left(g'_{\gamma(t)} g_{\gamma(t)}\right) = 2g'_{\gamma(t)} \frac{\partial g_{\gamma(t)}}{\partial \gamma'} \frac{d\gamma(t)}{dt} = 2g'_{\gamma(t)} V_{\gamma(t)} \frac{d\gamma(t)}{dt} \\ &= -g'_{\gamma(t)} V_{\gamma(t)} V_{\gamma(t)}^{-1} g_{\gamma(t)} = -g'_{\gamma(t)} g_{\gamma(t)} = -\left\|g_{\gamma(t)}\right\|^2 \end{aligned}$$

from which we can conclude that $\left\|g_{\gamma(t)}\right\|^2 = \|g_0\|^2 \exp(-t)$ .Since $\exp(-t) \to 0$ as $t \to \infty$, the solution $\gamma(t)$ to Equation (22) provides the path required to show that the moment conditions can be satisfied, provided $g_{\gamma(t)}$, $V_{\gamma(t)}$ and $V_{\gamma(t)}^{-1}$ exist at all $t \in \mathbb{R}^+$. Hence the existence of a suitable $\gamma(t)$ follows whenever we can establish the existence of $g_{\gamma(t)}$, $V_{\gamma(t)}$ and $V_{\gamma(t)}^{-1}$ for all $t \in \mathbb{R}^+$. As shown in Lemma A.1, for any $\gamma \in \mathbb{R}^{d_g}$, we have that g_γ, V_γ and V_γ^{-1} all exist. So the only possibility for the moment conditions to **not** be satisfied is to have $\gamma(t)$ diverging at some finite t. We now establish when this may or may not happen.

Letting $\gamma = r\eta$ for some unit vector η and applying Lemma A.2 for a fixed $z \in \mathcal{Z}_+$ with $X = \eta' g(U,z,\theta)$ and F equal to the distribution of $\eta' g(U,z,\theta)$ (which is nondegenerate since $\inf_{u\in\mathcal{U}} \eta' g(u,z,\theta) \neq \sup_{u\in\mathcal{U}} \eta' g(u,z,\theta)$ for $z \in \mathcal{Z}_+$ and $\operatorname{supp}\rho(\cdot|z) = \mathcal{U}$), implies that,

$$\eta' \tilde{g}(z,\theta,r\eta) = \frac{\int \eta' g(u,z,\theta) \exp(r\eta' g(u,z,\theta))\, d\rho(u|z)}{\int \exp(r\eta' g(u,z,\theta))\, d\rho(u|z)}$$

is less than $\sup_{u\in\mathcal{U}} \eta' g(u,z,\theta)$ for any finite r and only reaches it when $r \to \infty$. (For $z \in \mathcal{Z}\backslash\mathcal{Z}_+$, the quantity $\eta'\tilde{g}(z,\theta,r\eta)$ may be independent of γ for some value(s) of η, in which case the supremum is reached at all r, including when $r \to \infty$.) Next, consider the quantity $\eta' g_\gamma$ evaluated at $\gamma = r\eta$ in the limit as $r \to \infty$:

$$\lim_{r\to\infty} \eta' g_{r\eta} = \lim_{r\to\infty} \int \eta'\tilde{g}(z,\theta,r\eta)\, d\pi(z) = \int \lim_{r\to\infty} \eta'\tilde{g}(z,\theta,r\eta)\, d\pi(z),$$

where the interchange of the limit and the integral is justified by Lebesgue's monotone convergence theorem[19] since $\eta'\tilde{g}(z,\theta,r\eta)$ is monotone in r, as it can be readily verified that $\partial\eta'\tilde{g}(z,\theta,r\eta)/\partial r$ is equal to

$$\frac{\int (\eta' g(u,z,\theta) - \eta'\tilde{g}(z,\theta,r\eta))^2 \exp(r\eta' g(u,z,\theta))\, d\rho(u|z)}{\int \exp(r\eta' g(u,z,\theta))\, d\rho(u|z)} \geq 0,$$

where the interchange of derivatives and integration is allowed since the integrand is positive. It follows that, if $\|\gamma\| = r \to \infty$, not only does $\eta'\tilde{g}(z,\theta,r\eta)$ reach its maximum value at each z but so does $\eta' g_{r\eta}$. As this reasoning holds for any η, it follows that $g_{r\eta}$ would therefore converges to a point on the boundary of the convex set $\mathcal{K}_{\theta,\pi}$. Conversely, for finite r, and for all $z \in \mathcal{Z}_+$ (a set of positive probability under π), $\eta'\tilde{g}(z,\theta,r\eta)$ does not reach its maximum value and $\eta' g_\gamma$, the corresponding average over z, cannot reach its maximum value either. It follows that g_γ would lie in the interior of $\mathcal{K}_{\theta,\pi}$. Hence, $\|\gamma\| = r \to \infty$, iff g_γ converges to a point on the boundary of $\mathcal{K}_{\theta,\pi}$. Equivalently, Equation (22) only breaks down when g_γ reaches the boundary of $\mathcal{K}_{\theta,\pi}$.

Next, we note that if $g_{\gamma(t)}$ does not converge to the boundary of the convex set $\mathcal{K}_{\theta,\pi}$, it traces out (as t goes from 0 to infinity) a straight segment joining g_0 to 0 (see Figure 5a), because the change in $g_{\gamma(t)}$ is parallel to $g_{\gamma(t)}$ itself:

$$\frac{d}{dt} g_{\gamma(t)} = \frac{\partial g_{\gamma(t)}}{\partial \gamma'} \frac{d\gamma(t)}{dt} = -\frac{1}{2} V_{\gamma(t)} V_{\gamma(t)}^{-1} g_{\gamma(t)} = -\frac{1}{2} g_{\gamma(t)}.$$

However, if $g_{\gamma(t)}$ crossed the boundary of $\mathcal{K}_{\theta,\pi}$ somewhere along the segment from g_0 to 0 (see Figure 5b), the process would stop and g_γ could not reach 0. By definition,

[19] See Endou, Narita, and Shidama (2008), Section 52 for a statement of the Lebesgue Monotone Convergence Theorem generalized to extended reals (i.e. including "infinity"), thus allowing the interchange of integrals and limits for sequences of functions that have pointwise finite or infinite limits. Note that an infinite value of $\lim_{r\to\infty} \eta' g_{r\eta}$ is not pathological, as it only signifies that no inequality constraint is associated with the direction η. Also note that the Theorem's requirement that the integrand be nonnegative can be easily met by writing $\int \eta'\tilde{g}(z,\theta,r\eta)\, d\pi(z) = \int \eta' (\tilde{g}(z,\theta,r\eta) - \tilde{g}(z,\theta,r_0\eta))\, d\pi(z) + \int \eta'\tilde{g}(z,\theta,r_0\eta)\, d\pi(z)$ for all $r \geq r_0$ and for some $r_0 \in \mathbb{R}$.

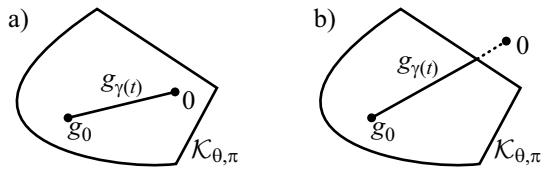

Figure 5: a) Path of $g_{\gamma(t)}$ when the origin is contained in $\mathcal{K}_{\theta,\pi}$. b) Path of $g_{\gamma(t)}$ when the origin is not contained in $\mathcal{K}_{\theta,\pi}$.

$g_0 \in \mathcal{K}_{\theta,\pi}$ (because $g_0 = E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right]$ for $\mu = \rho$). Since $\mathcal{K}_{\theta,\pi}$ is closed and convex, the segment joining g_0 to 0 is entirely contained in $\mathcal{K}_{\theta,\pi}$ iff 0 belongs to $\mathcal{K}_{\theta,\pi}$. It follows that $g_{\gamma(t)}$ cannot reach 0 if and only if 0 does not belong to $\mathcal{K}_{\theta,\pi}$. Since $\mathcal{K}_{\theta,\pi}$ is the closure of the set of all possible values of $E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right]$ for $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$, the process only fails if $\inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \left\|E_{\mu\times\pi}\left[g\left(U,Z,\theta\right)\right]\right\| \neq 0$. ■

Proof of Corollary 2.1. Apply Theorem 2.1 to the moment function $\mathring{g}\left(u,z,\mathring{\theta}\right) \equiv g\left(u,z,\theta\right) - \phi$ with $\mathring{\theta} \equiv (\theta,\phi)$. For any given θ, the identified set for ϕ gives the range of possible values of $E\left[g\left(u,z,\theta\right)\right]$. ■

Proof of Theorem 2.2. Let $\mathcal{K}_\theta = \text{Closure}\left\{E_{\mu\times\pi_0}\left[g\left(U,Z,\theta\right)\right] : \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}\right\}$, the set of all possible values of the moment conditions and note that, by definition $0 \in \mathcal{K}_\theta$ iff $\theta \in \Theta_0$. As in the proof of Theorem 2.1, $\mathcal{K}_\theta$ is convex. Hence, the set $\mathcal{K}_\theta$ can be written as an intersection of half spaces (Rockafellar (1970)):

$$\mathcal{K}_\theta = \cap_{\eta\in\delta\mathcal{B}_1}\left\{\kappa : \eta'\kappa \leq \bar{t}\left(\theta,\eta\right)\right\}, \tag{23}$$

where $\bar{t}\left(\theta,\eta\right)$ is a scalar function (the so-called "support function" of the set $\mathcal{K}_\theta$) that we will now determine through $\bar{t}\left(\theta,\eta\right) = \sup_{\kappa\in\mathcal{K}_\theta}\eta'\kappa$. Note that we allow $\bar{t}\left(\theta,\eta\right)$ to take the value "infinity" for some values of η, indicating that no constraint is associated with those values of η. (This convention differs from the one in Rockafellar (1970), where the domain of the support function is instead restricted so that $\bar{t}\left(\theta,\eta\right)$ is never infinite. These two conventions merely represent two different ways to state the same fact. Note that, without loss of generality, η can be restricted to the unit hypersphere, since both sides of an inequality can be scaled by a strictly positive constant without changing the set of values of κ satisfying the inequality.) By Corollary 2.1, $\mathcal{K}_\theta$ is also equal to $\text{Closure}\left\{E_{\pi_0}\left[\tilde{g}\left(Z,\theta,\gamma\right)\right] : \gamma \in \mathbb{R}^{d_g}\right\}$. Therefore, $\bar{t}\left(\theta,\eta\right) = \sup_{\gamma\in\mathbb{R}^{d_g}} E_{\pi_0}\left[\eta'\tilde{g}\left(Z,\theta,\gamma\right)\right]$. Considering one value of z and applying Lemma A.2 for a fixed z with $X = \eta' g\left(U,z,\theta\right)$ and F equal to the distribution of $\eta' g\left(U,z,\theta\right)$

implies that

$$\sup_{\gamma\in\mathbb{R}^{d_g}} \eta'\tilde{g}\left(z,\theta,\gamma\right) \leq \lim_{r\to\infty} \eta'\tilde{g}\left(z,\theta,r\eta\right) = \sup_{u\in\mathcal{U}} \eta' g\left(u,z,\theta\right).$$

Note that this indicates that the supremum of interest for a given η can be calculated using the same γ (or sequence of γ) at each value of z. Since multiplication by positive quantities and integration preserve inequalities, we also have

$$E\left[\eta'\tilde{g}\left(Z,\theta,\gamma\right)\right] \leq E\left[t\left(Z,\theta,\eta\right)\right] \equiv \bar{t}\left(\theta,\eta\right)$$

with $t\left(z,\theta,\eta\right) \equiv \lim_{r\to\infty} \eta'\tilde{g}\left(z,\theta,r\eta\right)$ or $t\left(z,\theta,\eta\right) = \sup_{u\in\mathcal{U}} \eta' g\left(u,z,\theta\right)$. We then need to verify if $0 \in \mathcal{K}_\theta$ using (23), which can be done by checking whether $\eta' 0 \leq \bar{t}\left(\theta,\eta\right)$ for all $\eta \in \delta\mathcal{B}_1$ or, equivalently $\bar{t}\left(\theta,\eta\right) = E\left[t\left(Z,\theta,\eta\right)\right] \geq 0$. ■

Proof of Theorem 4.1. By Theorem 2.1, at each finite J, it is clear that

$$\Theta_0^{(J)} = \left\{\theta\in\Theta : \inf_{\nu^{(J)}\in\mathcal{V}^{(J)}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \sup_{j\in\{0,\ldots,J\}} \left|E_{\mu\times\pi_0}\left[g_j\left(U,Z,\theta,\nu\right)\right]\right| = 0\right\}.$$

The sup is nondecreasing in J while the inf over $\nu^{(J)} \in \mathcal{V}^{(J)}$ is the same as over $\nu \in \mathcal{V}$. It follows that (i) $\Theta_0^{(J+1)} \subseteq \Theta_0^{(J)}$.

If $\theta \notin \Theta_0$, then $\inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \sup_{j\in\mathbb{N}^*} \left\|\int\int g_j\left(u,z,\theta,\nu\right) d\mu\left(u|z\right) d\pi\left(z\right)\right\| \neq 0$ and there exists a J_0 such that for all $J \geq J_0$ $\inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \sup_{j\in\{0,\ldots,J\}} \| \int\int g_j\left(u,z,\theta,\nu\right)$ $d\mu\left(u|z\right) d\pi\left(z\right) \| \neq 0$. If $\theta \in \Theta_0$ then $\inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \left\|\int\int g_j\left(u,z,\theta,\nu\right) d\mu\left(u|z\right) d\pi\left(z\right)\right\| =$ 0 for all $j \in \mathbb{N}^*$. It follows that (ii) $\cap_{J\in\mathbb{N}^*}\Theta_0^{(J)} = \Theta_0$.

Finally, $d_H\left(\Theta_0,\Theta^{(J+1)}\right) = d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J+1)}\right)$ since closure does not affect the Hausdorff distance. Also, $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J+1)}\right) = \sup_{\theta\in\bar{\Theta}^{(J+1)}} \inf_{\tilde{\theta}\in\bar{\Theta}_0} \left\|\theta-\tilde{\theta}\right\|$ because $\sup_{\theta\in\bar{\Theta}_0}$ $\inf_{\tilde{\theta}\in\bar{\Theta}^{(J+1)}} \left\|\theta-\tilde{\theta}\right\| = 0$ since $\bar{\Theta}_0 \subset \bar{\Theta}^{(J+1)}$. Next, $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J+1)}\right) \leq \sup_{\theta\in\bar{\Theta}^{(J)}} \inf_{\tilde{\theta}\in\bar{\Theta}_0} \left\|\theta-\tilde{\theta}\right\| =$ $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J)}\right)$ since $\bar{\Theta}^{(J+1)} \subseteq \bar{\Theta}^{(J)}$. Since $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J)}\right)$ forms a nonincreasing sequence and $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J)}\right) \geq 0$, we have $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J)}\right) \to c \geq 0$. We now show that c must be 0. We must have $\sup_{\theta\in\bar{\Theta}^{(J)}} \inf_{\tilde{\theta}\in\bar{\Theta}_0} \left\|\theta-\tilde{\theta}\right\| \geq c$ for all J. Since $\bar{\Theta}^{(J)}$ and $\bar{\Theta}_0$ are compact and the norm $\|\cdot\|$ is continuous, there exists $\theta_J \in \bar{\Theta}^{(J)}$ and $\tilde{\theta}_J \in \bar{\Theta}_0$ such that $\left\|\theta_J-\tilde{\theta}_J\right\| \geq c$ for all J. Since $\bar{\Theta}$ is compact, there exists a subsequence J_j such that θ_{J_j} and $\tilde{\theta}_{J_j}$ converge. Let $\theta_\infty = \lim_{j\to\infty}\theta_{J_j}$ and $\tilde{\theta}_\infty = \lim_{j\to\infty}\tilde{\theta}_{J_j}$. Note that $\theta_\infty \in \cap_{J\in\mathbb{N}^*}\bar{\Theta}^{(J)}$ for otherwise, eventually θ_∞ would lie at a finite distance from $\bar{\Theta}^{(J)}$ and $\left\|\theta_{J_j}-\theta_\infty\right\| \not\to 0$. Therefore $\theta_\infty \in \bar{\Theta}_0$, as $\bar{\Theta}_0$ is closed. Also $\tilde{\theta}_\infty \in \bar{\Theta}_0$ by construction, as $\bar{\Theta}_0$ is closed. Since $\tilde{\theta}_\infty$ minimizes the distance to θ_∞ and both $\tilde{\theta}_\infty$ and θ_∞ belong to $\bar{\Theta}_0$, it follows that $\left\|\theta_\infty-\tilde{\theta}_\infty\right\| = 0$ and that $c = 0$. Hence $d_H\left(\bar{\Theta}_0,\bar{\Theta}^{(J)}\right) \to 0$. ■

References

Afriat, S. N. (1973): "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method," *International Economic Review*, 14, 460–472.

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.

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," Working Paper 1676, Cowles Foundation, Yale University.

Andrews, D. W. K., and X. Shi (2008): "Inference for Parameters Defined by Conditional Moment Inequalities," Working Paper, Yale University.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763–814.

Bierens, H. J. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.

Blundell, R., M. Browning, and I. Crawford (2005): "Best Nonparametric Bounds on Demand Responses," Working Paper CWP12/05, cemmap.

Bontemps, C., T. Magnac, and E. Maurin (2007): "Set Identified Linear Models," Working Paper, Toulouse School of Economics.

Boyd, S., and L. Vandenberghe (2004): *Convex Optimization*. Cambridge University Press, New York.

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

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

CHAMBERLAIN, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets In Econometric Models," *Econometrica*, 75, 1243–1284.

CHERNOZHUKOV, V., E. KOCATULUM, AND K. MENZEL (2012): "Inference on Sets in Finance," Working Paper, Department of Economics, Massachusetts Institute of Technology.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2009): "Intersection bounds: estimation and inference," Working Paper CWP19/09, cemmap.

CHIBURIS, R. C. (2008): "Approximately Most Powerful Tests for Moment Inequalities," Working Paper, Princeton University.

CRAGG, J. C. (1997): "Using higher moments to estimate the simple errors-in-variables model," *Rand Journal of Economics*, 28, S71–S91.

CRESSIE, N., AND T. R. C. READ (1984): "Multinomial Goodness-of-Fit Tests," *J. R. Statist. Soc. B*, 46, 440–464.

CSISZAR, I. (1975): "I-Divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, 3, 146–158.

——— (1991): "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *Annals of Statistics*, 19, 2032–2066.

DONALD, S. G., G. IMBENS, AND W. NEWEY (2008): "Choosing the Number of Moments in Conditional Moment Restriction Models," Working Paper, Department of Economics, Massachusetts Institute of Technology.

DUDLEY, R. (2002): *Real Analysis and Probability*. Cambridge University Press, New York.

Ekeland, I., A. Galichon, and M. Henry (2010): "Optimal transportation and the falsifiability of incompletely specified economic models," *Economic Theory*, 42, 355–374.

Endou, N., K. Narita, and Y. Shidama (2008): "The Lebesgue Monotone Convergence Theorem," *Formalized Mathematics*, 16, 167–175.

Galichon, A., and M. Henry (2006): "Dilation Bootstrap: A methodology for constructing confidence regions with partially identified models," Working Paper, Harvard University and Columbia University.

Geweke, J., and M. Keane (2001): "Computationally Intensive Methods for Integration in Econometrics," in *Handbook of Econometrics*, vol. V. Elsevier Science.

Golan, A., G. Judge, and D. Miller (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley and Sons, New York.

Gourieroux, C., and A. Monfort (1997): *Simulation-Based Econometric Methods*. Oxford University Press, New York.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Haile, P. A., and E. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111, 1–51.

Hajivassiliou, V. A., and P. A. Ruud (1994): "Classical Estimation Methods for LDV Models Using Simulation," in *Handbook of Econometrics*, vol. IV, pp. 2384–2438. Elsevier Science.

Hansen, L. P. (1982): "Large sample properties of generalized method of moment estimators," *Econometrica*, 50, 1029–1054.

Harvey, A. (2004): "Forecasting with Unobserved Components Time Series Models," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann. Elsevier, North-Holland.

Harvey, A., S. J. Koopman, and N. Shephard (2004): *State Space and Unobserved Component Models: Theory and Applications*. Cambridge University Press, UK.

HAUSMAN, J., W. NEWEY, AND J. POWELL (1995): "Nonlinear Errors in Variables. Estimation of Some Engel Curves," *Journal of Econometrics*, 65, 205–233.

IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

KIM, K. (2008): "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," Working Paper, University of Minnesota.

KIRKPATRICK, S., C. D. GELATT, AND M. P. VECCHI (1983): "Optimization by Simulated Annealing," *Science*, 220, 671–680.

KITAMURA, Y. (2001): "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661–1672.

KITAMURA, Y., A. SANTOS, AND A. M. SHAIKH (2010): "On the Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," Working Paper, Department of Economics, University of Chicago.

KITAMURA, Y., AND M. STUTZER (1997): "An Information-Theoretic Alternative to Generalized Method of Moment Estimation," *Econometrica*, 65, 861–874.

KLEPPER, S., AND E. E. LEAMER (1984): "Consistent Sets of Estimates for Regressions with Errors in all Variables," *Econometrica*, 52, 163–183.

KULLBACK, S. (1959): *Information Theory and Statistics*. Wiley, Newyork.

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

MAGNAC, T., AND E. MAURIN (2008): "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data," *Review of Economic Studies*, 75, 835–864.

MANSKI, C. (1995): *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA.

——— (2003): *Partial Identification of Probability Distributions*. Springer-Verlag, New York.

MANSKI, C. F., AND E. TAMER (2002): "Inference on Regressions with Interval data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995–1026.

McFadden, D. L. (2005): "Revealed stochastic preference: a synthesis," *Economic Theory*, 26, 245–264.

Menzel, K. (2008): "Estimation And Inference With Many Moment Inequalities," Working Paper, MIT.

Molinari, F. (2008): "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 144, 81–117.

Newey, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engel, and D. L. McFadden, vol. IV. Elsevier Science.

Newey, W., and R. J. Smith (2004): "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Owen, A. B. (1988): "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.

——— (1990): "Empirical Likelihood Ratio Confidence Regions," *Annals of Statistics*, 18, 90–120.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2005): "Moment Inequalities and Their Application," Working Paper, Harvard University.

Ponomareva, M., and E. Tamer (2010): "Misspecication in Moment Inequality Models: Back to Moment Equalities?," *Econometrics Journal*, 10, forthcoming.

Qin, J., and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–325.

Rockafellar, R. T. (1970): *Convex Analysis*. Princeton University Press, Princeton.

ROEHRIG, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447.

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

ROSEN, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146, 107–117.

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

——— (2007b): "Point Estimation with Exponentially Tilted Empirical Likelihood," *Annals of Statistics*, 35, 634–672.

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.

SHEN, X. (1997): "On Methods of Sieves and Penalization," *Annals of Statistics*, 25, 2555–2591.

SHEN, X., J. SHI, AND W. H. WONG (1999): "Random Sieve Likelihood and General Regression Models," *Journal of the American Statistical Association*, 94(447), 835–846.

SHORE, J., AND R. JOHNSON (1980): "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions on Information Theory*, 26, 26–37.

STINCHCOMBE, M. B., AND H. WHITE (1998): "Consistent specification testing with nuisance parameters present only under the alternative," *Econometric Theory*, 14, 295–325.

VARIAN, H. R. (1982): "The Nonparametric Approach to Demand Analysis," *Econometrica*, 50, 945–973.

ZELLNER, A. (1997): "The Bayesian Method of Moments (BMOM)," *Advances in Econometrics*, 12, 85–105.

Supplementary Material for "Entropic Latent Variable Integration via Simulation"

Susanne M. Schennach
Department of Economics
Brown University
smschenn@brown.edu

This version: June 24, 2013

Abstract

This Supplementary 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.

B Introduction

This Supplementary 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.

C 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 $z \in \mathcal{Z} \backslash \mathcal{Z}'$, since such events have probability 0 and will not affect any unconditional probabilities or expectations. Also, recall that $g(u,z,\theta)$ is assumed 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 $C(z,\theta)\exp\left(-\|g(u,z,\theta) - g(\dot{u}(z),z,\theta)\|^2\right)$. 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 with expectation performed below are allowed. We then have

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

where

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

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

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

where we have (i) used the Cauchy-Schwartz inequality, (ii) used the fact that $\left(g_i(u,z,\theta)-\tilde{g}_i(z,\theta,\gamma)\right)^2 \leq \left\|g(u,z,\theta)-\tilde{g}(z,\theta,\gamma)\right\|^2$ for $i=1,\ldots,d_g$ and (iii) multiplied the numerator and denominator by the same non-vanishing factor $\exp\left(-\gamma' g\left(\dot{u}(z),z,\theta\right)\right)$.

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

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

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

$$\begin{aligned}
&E_\rho\left[\left\|g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right\|^2\exp\left(\gamma'\left(g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right)\right)|Z=z\right] \\
\leq\ & \sum_{j=0}^{d_g} A_j E_\rho\left[\exp\left(\gamma_j'\left(g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right)\right)|Z=z\right] \qquad (3)
\end{aligned}$$

for some finite $A_0,\ldots,A_{d_g} \in \mathbb{R}^+$ and some $\gamma_0,\ldots,\gamma_{d_g}$, 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 (3) we have, by the definition of ρ,

$$\begin{aligned}
&A_j E_\rho\left[\exp\left(\gamma_j'\left(g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right)\right)|Z=z\right] \\
=\ & A_j\frac{E_\lambda\left[\exp\left(\gamma_j'\left(g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right)-\left\|g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right\|^2\right)|Z=z\right]}{E_\lambda\left[\exp\left(-\left\|g(U,Z,\theta)-g\left(\dot{u}(z),z,\theta\right)\right\|^2\right)|Z=z\right]}, \qquad (4)
\end{aligned}$$

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

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

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 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 (4) as

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

where we have used (i) Inequality (5) (ii) the fact that $\sup_{x\in\mathbb{R}}\left(\left\|\gamma_j\right\|x-x^2\right)=\left\|\gamma_j\right\|^2/4$ 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 (2):

$$\begin{aligned}
& E_\rho\left[\exp\left(\gamma'\left(g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right)\right)|Z=z\right] \\
=\ & \frac{E_\lambda\left[\exp\left(\gamma'\left(g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right)\right)\exp\left(-\left\|g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right\|^2\right)|Z=z\right]}{E_\lambda\left[\exp\left(-\left\|g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right\|^2\right)|Z=z\right]} \\
\geq\ & E_\lambda\left[\exp\left(\gamma'\left(g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right)\right)\exp\left(-\left\|g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right\|^2\right)|Z=z\right] \\
\geq\ & E_\lambda\left[\exp\left(\gamma'\left(g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right)\right)\exp\left(-\left\|g(U,Z,\theta)-g(\dot{u}(z),z,\theta)\right\|^2\right)\times\right. \\
& \left.1\left(g(U,Z,\theta)=g(\dot{u}(z),z,\theta)\right)|Z=z\right] \\
=\ & E_\lambda\left[1\left(g(U,Z,\theta)=g(\dot{u}(z),z,\theta)\right)|Z=z\right] \\
\geq\ & E_\lambda\left[1\left(U=\dot{u}\right)|Z=z\right]=q, \qquad (7)
\end{aligned}$$

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

$$E_\pi\left[\frac{E_\rho\left[\left\|g(U,Z,\theta) - \tilde{g}(Z,\theta,\gamma)\right\|^2 \exp\left(\gamma' g(U,Z,\theta)\right)|Z\right]}{E_\rho\left[\exp\left(\gamma' g(U,Z,\theta)\right)|Z\right]}\right] \leq \bar{A}\exp\left(\left(\|\gamma\| + \varepsilon\right)^2/4\right), \tag{8}$$

where $\bar{A} \equiv q^{-2}\sum_{j=0}^{d_g} A_j$. This bound on the second derivative of $E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g(U,Z,\theta)\right)|Z\right]\right] \equiv \tilde{M}(\gamma)$ 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:

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

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

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

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

$$
\begin{aligned}
& E_\pi\left[\frac{E_\lambda\left[\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|\exp\left(-\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|^2\right)|Z\right]}{E_\lambda\left[\exp\left(-\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|^2\right)|Z\right]}\right]+ \\
& +E_\pi\left[E_\lambda\left[\|g(\dot{u}(Z),Z,\theta)\|\right]\right] \\
\leq\ & E_\pi\left[q^{-1}E_\lambda\left[\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|\exp\left(-\|g(U,Z,\theta)-g(\dot{u}(Z),Z,\theta)\|^2\right)|Z\right]\right]+ \\
& +E_\pi\left[\|g(\dot{u}(Z),Z,\theta)\|\right] \\
\leq\ & E_\pi\left[q^{-1}E_\lambda\left[\left(\sup_{x\in\mathbb{R}}|x|\exp\left(-x^2\right)\right)|Z\right]\right]+E_\pi\left[\|g(\dot{u}(Z),Z,\theta)\|\right] \\
\leq\ & E_\pi\left[q^{-1}E_\lambda\left[1|Z\right]\right]+E_\pi\left[\|g(\dot{u}(Z),Z,\theta)\|\right]=q^{-1}+E_\pi\left[\|g(\dot{u}(Z),Z,\theta)\|\right] \\
\leq\ & q^{-1}+E_\pi\left[\inf_{u\in\mathcal{U}}\|g(u,Z,\theta)\|\right]+\omega \\
\leq\ & q^{-1}+E_{\mu\times\pi}\left[\|g(U,Z,\theta)\|\right]+\omega<\infty, \qquad (10)
\end{aligned}
$$

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

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

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 $E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g(U,Z,\theta)\right)|Z\right]\right]$ satisfies the existence and differentiability conditions of Definition 2.2. ■

Proof of Lemma A.1. If $E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g(U,Z,\theta)\right)|Z\right]\right]$ exists for all $\gamma\in\mathbb{R}^{d_g}$, then $E_\rho\left[\exp\left(\gamma' g(U,Z,\theta)\right)|Z=z\right]$ 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 zero 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

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

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

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

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

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

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 $g_\gamma=\frac{\partial}{\partial\gamma}E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g\left(U,Z,\theta\right)\right)|Z\right]\right]$ and $V_\gamma=\frac{\partial^2}{\partial\gamma\partial\gamma'}E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g\left(U,Z,\theta\right)\right)|Z\right]\right]$, which both exist because $E_\pi\left[\ln E_\rho\left[\exp\left(\gamma' g\left(U,Z,\theta\right)\right)|Z\right]\right]$ 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 $\eta' g\left(U,z,\theta\right)$ (conditional on z) under the measure $\tilde{\rho}\left(u|z\right)$ defined via

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

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

Proposition C.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 corresponding unconditional expectation $E[Y]$) are well-defined,*[1] *then the restriction $E[Y|X]=0$*

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

(with probability 1 *under the distribution of* X*) is equivalent to a countable set of unconditional moment restrictions.*

Proof of Proposition C.1. 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 $E\left[Ye^{\mathbf{i}\xi'X}\right]$, 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 $E\left[Ye^{\mathbf{i}\xi'X}\right]$ is continuous in ξ. Hence, having $E\left[Ye^{\mathbf{i}\xi'X}\right] = 0$ for all rational ξ implies that $E\left[Ye^{\mathbf{i}\xi'X}\right] = 0$ for all $\xi \in \mathbb{R}$. The inverse Fourier transform of $E\left[Ye^{\mathbf{i}\xi'X}\right] = E\left[E[Y|X]e^{\mathbf{i}\xi'X}\right]$ therefore vanishes almost everywhere. Since $E\left[e^{\mathbf{i}\xi'X}E[Y|X]\right] = \int e^{\mathbf{i}\xi'X}E[Y|X]\,dF(x)$, 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] ($E\left[Ye^{\mathbf{i}\xi'X}\right] = 0$ 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). ■

Proposition C.2 *Independence restrictions can be imposed via a countable number of moment factorization restriction of the form (17), i.e. without loss of generality, the index* t *can be discrete.*

Proof of Proposition C.2. 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 iff

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

for all $\xi, \eta \in \mathbb{R}$ where $\mathbf{i} = \sqrt{-1}$. By result 13.4-A in Loève (1977), all three expectations in (11) are continuous functions of ξ and η. Hence, imposing the constraint (11) at all rational ξ and η is sufficient to imply that (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. ■

[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.

D Additional simulation examples

D.1 Regression with interval-valued data

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

$$\begin{aligned} Y^* &= X\theta_1 + V \text{ with } \theta_1 = 1 \\ \overline{Y} &= \lceil Y^* \rceil \\ \underline{Y} &= \lfloor Y^* \rfloor . \end{aligned}$$

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 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 Section I of the Supplementary Material. (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 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 giving 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 ensuring that the variance of the residuals (conditional on X) is not correlated with X^2. The moment function would then be*

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

where $V = \underline{Y} + U\left(\overline{Y} - \underline{Y}\right) - X\theta_1$ 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 properly handling the sampling noise.

[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.

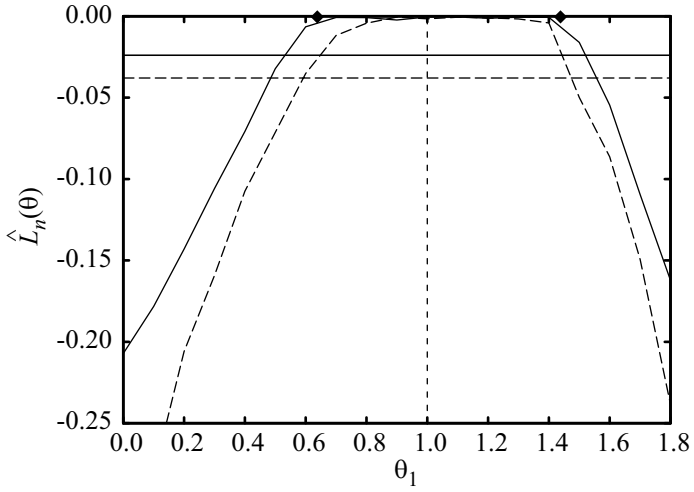

Figure 1: Objective function for an interval-valued data regression model. The upper diamonds mark the standard bounds for this model while the true value of the parameter is indicated by a vertical dashed line. The solid curve is obtained with the usual uncorrelatedness assumption while the dashed line is for a model also assuming 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.

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

D.2 Censored regression

We now apply our method to the censored regression of Example 1.2 by generating an iid sample of 250 observations as follows:

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

The algorithm of Section 2.3 (with empirical likelihood) was used with $R = 900$, after 100 equilibration steps. Figure 2a) 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 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^* give 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 the 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 2b).

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:*

$$g\left(U,Z,\theta\right)=\left[\begin{array}{c}\left(Y+U\mathbf{1}\left(Y=c\right)-X\theta\right)X\\ \left(Y+U\mathbf{1}\left(Y=c\right)-X\theta\right)^{2}X\end{array}\right].$$

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.

D.3 Nonlinear errors-in-variable model without side information

We now consider a model for which a pre-existing analysis of identification is not available.

Example 1.5 (continued) *Consider a nonlinear errors-in-variables model*

$$\begin{array}{rcl}Y & = & r\left(X^{*},\theta\right)+V_{2}\\ X & = & X^{*}+V_{1},\end{array}\tag{13}$$

where $r\left(X^{*},\theta\right)$ is a given parametric specification with unknown parameter vector $\theta=(\theta_1,\theta_2)$ and we impose the following vector of moment conditions: $g\left(U,Z,\theta\right)=$

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

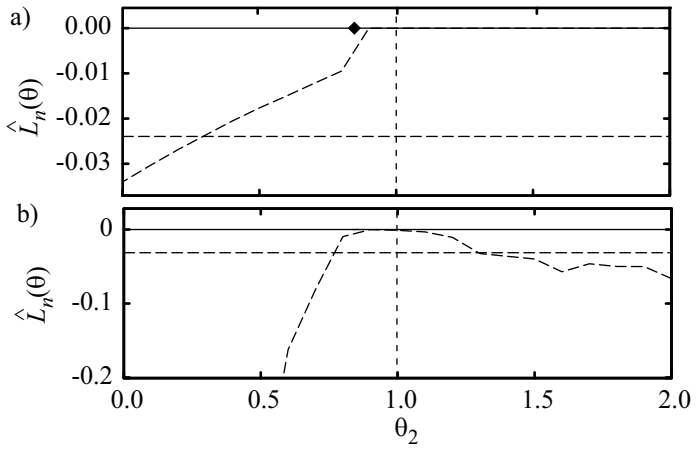

Figure 2: Objective function for a censored regression model. a) Result obtained with the usual uncorrelatedness and zero mean assumptions on the residuals. The upper diamond mark the well-known lower bound for this model. b) Same exercise while assuming, in addition, that the variance of the residuals 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.

$(V_1, V_2, V_1 \partial r(X^*, \theta)/\partial\theta_1, V_1 \partial r(X^*, \theta)/\partial\theta_2, V_2 \partial r(X^*, \theta)/\partial\theta_1, V_2 \partial r(X^*, \theta)/\partial\theta_2, V_1 V_2)'$. *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 iid observation is generated according to Equation (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 3 for two specifications:

$$r(X^*, \theta) = \theta_1 X^* + \theta_2 (X^*)^2 \quad (14)$$

$$r(X^*, \theta) = \theta_1 X^* + \theta_2 \exp(X^*). \quad (15)$$

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 (15). In contrast, our method applies directly — only trivial changes in the program handling the standard measurement error problem were needed.

The time need to complete these simulations range from a few minutes (for the simplest models) to a few hours (for the one with 27 moment conditions) on an average single processor personal computer in 2008-2009 and using the Gauss language.

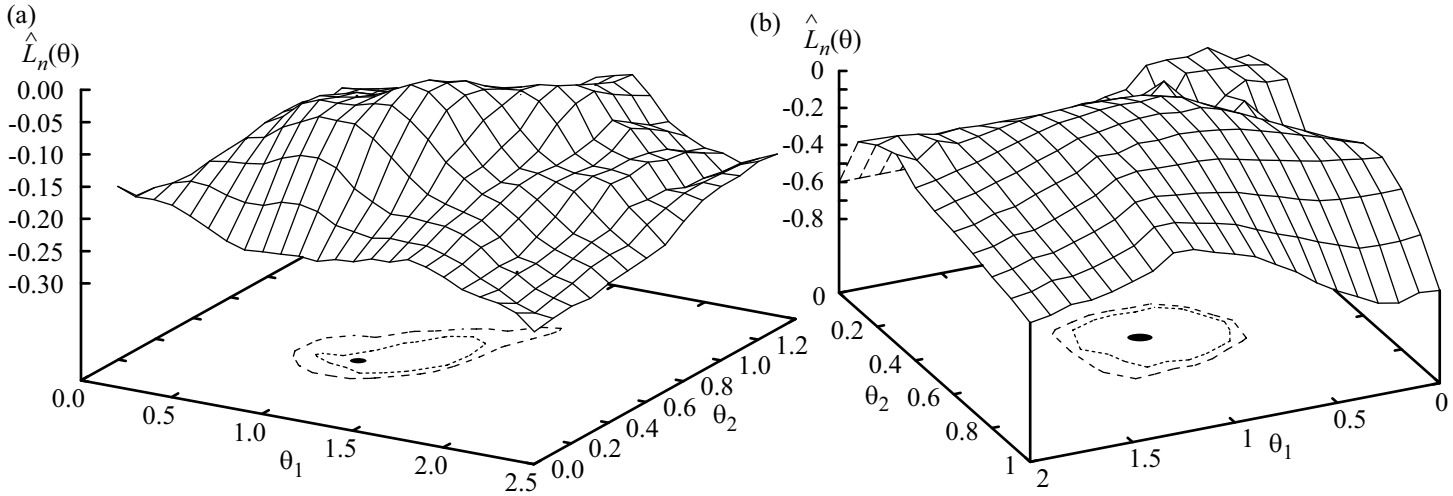

Figure 3: Objective function for (a) the polynomial measurement error model of Equation (14) and (b) for the linear-exponential measurement error model of Equation (15). The base plane shows the joint critical region at the 95% and the 99% levels while the true value of the parameters is indicated by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

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 (regarless 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).

E Extended notion of the Identified Set

E.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 $E_{\mu_k \times \pi}\left[g\left(U, Z, \theta\right)\right] \to 0$. The set Θ_0 (from Equation (3)) is preferable to Θ_0^* (Equation (2)) for two reasons:

1. 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 cause 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 arbitrarily close to that which *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.

2. 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 Section E.2 below. This invariance is important because the choice of 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 (2006) make uniform integrability assumptions to rule out improper distributions). But this is unnecessary here, since, in light of point 1 above, the distinction between Θ_0 and Θ_0^* is inconsequential in practice, as it could never be detected and since Point 2 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.

E.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 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.

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 $\Theta_0 = \{0\}$.

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

   (a) We then have that $\Theta_0^* = \{0\}$ because $E_\mu\left[\tilde{g}\left(\widetilde{U}, Z, \theta\right)\right] = 0$ 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 case, Θ_0^* is not parametrization invariant, while Θ_0 is.

F 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:

$$\int\int \ln\left(f\left(u|z\right)\right) d\rho\left(u|z\right) d\pi\left(z\right) - \gamma' \int\int g\left(u, z, \theta\right) f\left(u|z\right) d\rho\left(u|z\right) d\pi\left(z\right)$$
$$- \int \phi\left(z\right)\left(\int f\left(u|z\right) d\rho\left(u|z\right) - 1\right) d\pi\left(z\right).$$

The first order condition is then:

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

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

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

or, after rearranging,

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

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

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

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 (16) as

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

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

G Inference methods

As models defined via moment conditions involving 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.

G.1 Objective functions and confidence regions

We first introduce a general class of possible objective functions.

Definition G.1 *Given an iid sample $Z_1, \ldots, Z_n$, we consider an empirical objective function admitting the representation*

$$\begin{aligned} \hat{L}_n(\theta) &= \sup_{\gamma \in \mathbb{R}^{d_g}} \hat{L}_n(\theta,\gamma) \\ \hat{L}_n(\theta,\gamma) &= -\frac{1}{2} \hat{g}'(\theta,\gamma) W(\theta,\gamma) \hat{g}(\theta,\gamma) + \hat{R}_n(\theta,\gamma), \end{aligned}$$

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

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

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

and where the remainder satisfies

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

and is such that

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

for some $C>0$ w.p.a. 1. We also assume throughout that the $\rho\left(u|z\right)$ used to construct $\tilde{g}\left(Z,\theta,\gamma\right)$ is as in Definition 2.2 and that the unobservables U_i are iid.

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

$$\begin{aligned}\hat{L}_n^{EL}\left(\theta\right) &= \sup_{\gamma\in\mathbb{R}^{d_g}}\inf_{\lambda\in\mathbb{R}^{d_g}}\frac{1}{n}\sum_{i=1}^{n}-\ln\left(1-\lambda'\tilde{g}\left(Z_i,\theta,\gamma\right)\right)\\ \hat{L}_n^{CUE}\left(\theta\right) &= \sup_{\gamma\in\mathbb{R}^{d_g}}-\frac{1}{2}\hat{g}'\left(\theta,\gamma\right)\hat{V}^{-1}\left(\theta,\gamma\right)\hat{g}\left(\theta,\gamma\right)\text{ with }\hat{V}\left(\theta,\gamma\right)=\frac{1}{n}\sum_{i=1}^{n}\tilde{g}\left(Z_i,\theta,\gamma\right)\tilde{g}'\left(Z_i,\theta,\gamma\right).\end{aligned}$$

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

It should be noted that, although $\hat{L}_n\left(\theta,\gamma\right)$ depends on the choice of ρ in Definition 2.2, the objective function $\hat{L}_n\left(\theta\right)$ 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 $\mathcal{B}_0\equiv\left\{\beta\in\mathcal{B}:\inf_{\eta\in\mathcal{N}_\beta}\inf_{\gamma\in\mathbb{R}^{d_g}}\left\|E\left[\tilde{g}\left(Z_i,\theta,\gamma\right)\right]\right\|=0\right\}$ via the "profiled" statistic:

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

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 $\hat{Q}_n\left(\beta\right)$ 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

$$\hat{\mathcal{B}} = \left\{ \beta : n\hat{Q}_n(\beta) \leq \hat{c}_\alpha \right\}, \tag{20}$$

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

G.2 Consistency

Consistency of $\hat{\mathcal{B}}$ (in the sense that the Hausdorff distance between $\hat{\mathcal{B}}$ and $\mathcal{B}_0$ goes to zero 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 demanding 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 $\bar{g}(\theta,\gamma) = E[\tilde{g}(Z_i,\theta,\gamma)]$, $\mathcal{K}_\theta \equiv \text{Closure}\left\{ E_{\pi_0}[\tilde{g}(Z_i,\theta,\gamma)] : \gamma \in \mathbb{R}^{d_g} \right\}$ and $\mathcal{K}_\theta^* = \mathcal{K}_\theta \cap \mathcal{C}$, where $\mathcal{C}$ is a sufficiently large compact convex set containing a neighborhood of $\{0\}$. The reparametrized moment functions are then

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

with γ_j $(j=1,2,\ldots)$ such that $\bar{g}\left(\theta,\gamma_j\right) = \kappa + (\bar{\kappa}-\kappa)/j$, 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 (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, sometimes, itself 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 making the sample moment as small as possible, i.e. values of κ near zero. The constraint $\kappa \in \mathcal{C}$ is therefore not binding with probability approaching one. The reparametrized moment functions $\tilde{g}_\kappa(\cdot,\theta,\kappa)$ are then indexed over a domain $\cup_{\theta\in\Theta}\{\theta\} \times \mathcal{K}_\theta^*$, which is compact by construction.

[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.

Remark G.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 implemention 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)), i.e., their normalized sample averages converge to a tight Gaussian process in the sup metric.[8] A sufficient condition is as follows:

Assumption G.1 *Z_i is iid, $E\left[\left\|\tilde{g}\left(Z_i,\tilde{\theta},\tilde{\gamma}\right)\right\|^2\right]<\infty$ 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 positive δ less than some $\delta_0\in\,]0,\infty[$,*

$$\sup_{\theta_1\in\Theta}\sup_{\gamma_1\in\mathbb{R}^{d_g}}E\left[\sup_{\gamma_2\in\mathbb{R}^{d_g}:\|\bar{g}(\theta_1,\gamma_2)-\bar{g}(\theta_1,\gamma_1)\|\leq\delta}\|\tilde{g}(Z_i,\theta_1,\gamma_2)-\tilde{g}(Z_i,\theta_1,\gamma_1)\|^2\right] = O\left(\delta^{\alpha}\right) \quad (22)$$

$$\sup_{\theta_1\in\Theta}\sup_{\gamma_1\in\mathbb{R}^{d_g}}E\left[\sup_{\theta_2\in\Theta:\|\theta_2-\theta_1\|\leq\delta}\|\tilde{g}(Z_i,\theta_2,\gamma_1)-\tilde{g}(Z_i,\theta_1,\gamma_1)\|^2\right] = O\left(\delta^{\alpha}\right) \quad (23)$$

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 $\mathcal{K}_\theta^*$ being not piecewise-differentiable, a somewhat pathological setting. Note that the metric used for γ in (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 G.1 implies that $\sup_{\theta\in\Theta}\sup_{\gamma\in\mathbb{R}^{d_g}}\|\tilde{g}(Z_i,\theta,\gamma)\|=O_p\left(n^{-1/2}\right)$, 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 G.1 is implied by more primitive conditions on $g(u,z,\theta)$, such moment existence and smoothness. For instance, see Lemma G.1 in Section G.5 for the interval-valued data model of Example 1.1.

[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.

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 (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 $\hat{c}_\alpha \propto \ln n$).

G.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, e.g., in Imbens and Manski (2004) and Chernozhukov, Hong, and Tamer (2007), there are two main types of confidence region: Pointwise regions satisfying $\lim_{n\to\infty} P\left[\beta_0 \in \hat{\mathcal{B}}\right] \geq 1-\alpha$ for any $\beta_0 \in \mathcal{B}_0$ and "setwise" regions satisfying $\lim_{n\to\infty} P\left[\mathcal{B}_0 \subset \hat{\mathcal{B}}\right] \geq 1-\alpha$. 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 having a smaller confidence regions than a point-identified model nested as a special case of this family. Andrews

[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.

and Guggenberger (2009) provides 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 which may require special attention. Establishing the stochastic convergence of $\hat{L}_n(\theta, \gamma)$ can be accomplished as in the consistency result (see Assumption G.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]

Another technical issue is that the set over which the maximizations take place must be sufficiently regular, i.e. 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 Jia (2008)).

G.4 Simple but conservative critical values

A companion paper (Schennach (2009)) presents an asymptotic treatment providing 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

[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 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.

confidence regions of the form

$$\left\{\beta \in \mathcal{B} : \sup_{\eta \in \mathcal{N}_\beta} -n\hat{L}\left((\beta,\eta)\right) \leq \hat{c}_\alpha\right\}$$

but does not allow for a subtraction of the objective function of the "unrestricted model" as in (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 one, and confidence regions that are not conservative can be obtained in this fashion without recourse to resampling. This condition is closely related (though not identical) to the so-called "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 G.2 *Z_i is iid.*

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

Assumption G.4 *$E\left[\|\tilde{g}(Z,\theta,\gamma)\|^2\right] < \infty$ $\forall\theta \in \Theta$ and $\gamma \in \Gamma_\theta$*

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

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

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 $E\left[\tilde{g}(Z,\theta,\gamma)\right] = 0$ with $q = \text{rank}(E\left[\tilde{g}(Z,\theta,\gamma)\,\tilde{g}'(Z,\theta,\gamma)\right])$ for the Empirical Likelihood (EL) objective function. His proof first proceeds by showing that EL has the representation of Definition G.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 $\mathcal{R} = \left\{(\theta,\gamma) \in \Theta \times \mathbb{R}^{d_g} : -2\hat{L}_n(\theta,\gamma) \geq \chi^2_{d_g,\alpha}\right\}$

[11] Chernozhukov, Hong, and Tamer (2007) state 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.

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 $-2n\hat{L}_n(\theta) \geq \chi^2_{d_g,\alpha}$. ■

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.

G.5 Primitive conditions for Assumption G.1 in Example 1.1

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

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

$$\begin{aligned} \tilde{g}(z, \theta, \gamma) &= \frac{\int_0^1 \left(\underline{y} + u\left(\overline{y} - \underline{y}\right) - x\theta\right) x \exp\left(\gamma\left(\underline{y} + u\left(\overline{y} - \underline{y}\right) - x\theta\right) x\right) du}{\int_0^1 \exp\left(\gamma\left(\underline{y} + u\left(\overline{y} - \underline{y}\right) - x\theta\right) x\right) du} \\ &= \left(\underline{y} - \theta x\right) x + \frac{w}{1 - e^{-\gamma\left(\overline{y} - \underline{y}\right) x}} - \frac{1}{\gamma}, \end{aligned} \tag{24}$$

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

$$B_{\theta_1,\gamma_1} \equiv E\left[\sup_{\theta_2 \in \Theta : \|\theta_2 - \theta_1\| \leq \delta} \|\tilde{g}(Z, \theta_2, \gamma_1) - \tilde{g}(Z, \theta_1, \gamma_1)\|^2\right] \leq E\left[\sup_{\bar{\theta} \in \Theta : \left\|\bar{\theta} - \theta_1\right\| \leq \delta} \left\|\nabla_{\theta'} \tilde{g}_\gamma\left(Z, \bar{\theta}, \gamma_1\right)\right\|^2\right] \delta^2,$$

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

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

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

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

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

By a mean value argument, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 (30) as

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

(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 (30) as

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

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

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

which is $O\left(\delta^2\right)$ if $E[W^4] < \infty$. This establishes (22) in Assumption G.1. ■

H 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 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), Chap. 9.5.2) or quasi-Newton method, such as L-BGFS (Nocedal (1980)).

I 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 2.1 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, $g(u,z,\theta) = \left(\underline{y} + u\left(\overline{y}-\underline{y}\right) - \theta x\right)x$ with $z = \left(x,\overline{y},\underline{y}\right)$ 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 $\eta = \pm 1$:

$$\begin{aligned}\eta'\tilde{g}(z,\theta,\eta r) &= \frac{\int_0^1 \eta g(u,z,\theta)\exp(r\eta g(u,z,\theta))\,du}{\int_0^1 \exp(r\eta g(u,z,\theta))\,du} \\ &= \eta\left(\underline{y}-\theta x\right)x + \left[\eta b\frac{\left(1+\frac{1}{r\eta b}\left(e^{-r\eta b}-1\right)\right)}{1-e^{-r\eta b}}\right]_{b=\left(\overline{y}-\underline{y}\right)x}\end{aligned}$$

and therefore

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

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

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

For $\eta = -1$, we have

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

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

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

Isolating θ yields:

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

in agreement with, e.g., 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.

J 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 consisting 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 J.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* $\left[\underline{Y}_i, \overline{Y}_i\right] \times X_i$. *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 *corresponding to these weights is simply the slope of the regression of* $\left(\underline{Y}_i + \overline{Y}_i\right)/2$ *on* X_i, *because the uniform weights simply result in averaging values in the interval* $\left[\underline{Y}_i, \overline{Y}_i\right]$. *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* $\overline{Y}_i \mathbf{1}\left[X_i > 0\right] + \underline{Y}_i \mathbf{1}\left[X_i < 0\right]$ *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 zero). 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.

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," Working Paper 1676, Cowles Foundation, Yale University.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Boyd, S., and L. Vandenberghe (2004): *Convex Optimization*. Cambridge University Press, New York.

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

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

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

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

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets In Econometric Models," *Econometrica*, 75, 1243–1284.

Claeskens, G. (2004): "Restricted likelihood ratio lack-of-fit tests using mixed spline models," *Journal of the Royal Statistical Society B*, 66, 909–926.

Galichon, A., and M. Henry (2006): "Dilation Bootstrap: A methodology for constructing confidence regions with partially identified models," Working Paper, Harvard University and Columbia University.

Golan, A., G. Judge, and D. Miller (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley and Sons, New York.

Imbens, G., and C. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

KITAMURA, Y. (2001): "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661–1672.

KITAMURA, Y., A. SANTOS, AND A. M. SHAIKH (2010): "On the Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," Working Paper, Department of Economics, University of Chicago.

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

MANSKI, C. F., AND E. TAMER (2002): "Inference on Regressions with Interval data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

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

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

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

NOCEDAL, J. (1980): "Updating Quasi-Newton Matrices With Limited Storage," *Math. Comp.*, 35, 773–782.

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

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

ROSEN, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146, 107–117.

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

——— (2009): "Simple conservative confidence regions for a class of set identified models," Working Paper, University of Chicago.

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.

SHEN, X., J. SHI, AND W. H. WONG (1999): "Random Sieve Likelihood and General Regression Models," *Journal of the American Statistical Association*, 94(447), 835–846.

SILVAPULLE, M. J., AND P. L. SEN (2005): *Constrained Statistical Inference*. Wiley, New Jersey.

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

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes. With Applications to Statistics*. New York: Springer.