ENTROPIC LATENT VARIABLE INTEGRATION VIA SIMULATION

Author(s): Susanne M. Schennach

Source: *Econometrica*, January 2014, Vol. 82, No. 1 (January 2014), pp. 345-385

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/24029178

Econometrica, Vol. 82, No. 1 (January, 2014), 345–385

ENTROPIC LATENT VARIABLE INTEGRATION VIA SIMULATION

BY SUSANNE M. SCHENNACH[1]

This paper introduces a general method to convert a model defined by moment conditions that involve both observed and unobserved variables into equivalent moment conditions that involve 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 generalized method of moments-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.

1. INTRODUCTION

1.1. *Outline*

OUR GOAL IS TO FIND THE VALUE(S) of a parameter $\theta \in \mathbb{R}^{d_\theta}$ that satisfies 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 that enter the moment conditions are not observable. Specifically, the moment conditions have the general form

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

where g is a d_g-dimensional vector of nonlinear measurable functions that depend 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

[1]The author would like to thank Daniel Wilhelm and 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 she acknowledges support from the National Science Foundation via Grants SES-0752699 and SES-1061263/1156347.

 DOI: 10.3982/ECTA9748

identification problem. The use of a dual representation for this purpose was previously suggested in the important works of Galichon and Henry (2013) 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 that enable a simple nonparametric generalization of the method of simulated moments), and weakening the necessary regularity conditions (notably, allowing for unbounded moment functions $g(u, z, \theta)$, such as the mean of a variable with unbounded support).

In essence, the method consists of eliminating the unobservables by averaging the moment function $g(U, Z, \theta)$ 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 that involve 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 (generalized empirical likelihood (GEL) or exponentially tilted empirical likelihood; see, for instance, Owen (1990), Qin and Lawless (1994), Kitamura and Stutzer (1997), Imbens, Spady, and Johnson (1998), Newey and Smith (2004), Schennach (2007)). 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 that solve the averaged moment conditions are the same as the values that solve 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 Barwick (2012), 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 that motivate the usefulness of the class of models 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 Appendix B of the Supplemental Material (Schennach (2014)). The Supplemental 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 (e.g., Manski and Tamer (2002)): 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 $[\underline{Y}, \overline{Y}]$, which may vary across individuals. The scalar disturbance V satisfies $E[VX] = 0$. This model fits our framework with $Z = (\underline{Y}, \overline{Y}, X)'$, $U = (Y^* - \underline{Y})/(\overline{Y} - \underline{Y})$, $\mathcal{U} = [0, 1]$, and

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

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 the scalar dependent variable Y^* is not directly observable.[2] In-

[2]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 (2011), even in the presence of potential model misspecification.

stead, one observes

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

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

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

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.

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 recast 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 $c \in \mathcal{C} \setminus \{C\}$. The use of such a "revealed preference" argument has a 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 (2008), 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.[3] This model fits our framework with[4] $Z = (C, X', W')', \mathcal{U} = \mathbb{R}$, and

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

The second moment condition imposes that the fraction of the population that satisfies 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 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:

$$Y = X^*\theta + V_2,$$
$$X = X^* + V_1.$$

A natural set of moment conditions in this case could be

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

Even though a data set 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

$$V_2 = Y - X^*\theta,$$
$$V_1 = X - X^*.$$

[3]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.

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

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 leads 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 that take 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 to be 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, Chapter 10.2)) conditional probability measures[5] 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

[5]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.

probabilities under π. In an integral with respect to ν, the differential element $d\nu(s, c)$ is 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 Euclidean 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*[6] *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 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 is discussed in Section 4.1 below. When implementing the method, it is not necessary to transform the model into 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 with regard to 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.

[6]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.

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))[7]

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

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

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

As discussed in Appendix D of the Supplemental 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 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.[8] 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 the following statements hold:

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

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

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

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

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*[9]

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

where $C(z,\theta) = (E_{\lambda(\cdot|\cdot;\theta)}[\exp(-\|g(U,z,\theta) - g(\dot{u}(z,\theta),z,\theta)\|^2)|Z=z])^{-1}$, *is a normalization constant and* $\lambda(\cdot|\cdot;\theta)$ *is a conditional probability measure that satisfies* $\operatorname{supp}\lambda(\cdot|z;\theta) = \mathcal{U}$ *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 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* $\pi \in \mathcal{P}_{\mathcal{Z}}$,

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

if and only if[10]

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

where

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

[9]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, that is, $d\rho(\cdot|z)/d\lambda = a(\cdot,z)$."

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

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* $\pi \in \mathcal{P}_{\mathcal{Z}}$,

$$\begin{aligned}&\text{Closure}\left\{E_{\mu\times\pi}\left[g(U,Z,\theta)\right]:\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}\right\}\\&\quad=\text{Closure}\left\{E_{\pi}\left[\tilde{g}(Z,\theta,\gamma)\right]:\gamma\in\mathbb{R}^{d_{\tilde{g}}}\right\}.\end{aligned}$$

Theorem 2.1 proves that the infinite-dimensional problem of establishing the existence of some measure μ that solves 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 on 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 what is special about the exponential family selected is its "least favorable" property, that is, 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, Appendix H in the Supplemental 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 that 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 pseudo-likelihoods (see Definition 1 in Gourieroux, Monfort, and Trognon (1984)). There is an important difference, however: The choice of the family used in pseudo-likelihoods 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 $\{\hat{g}(\theta, \gamma) : \gamma \in \mathbb{R}^{d_g}\}$ 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 $\mathcal{U}$ so as 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 that solve 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 $E_\pi[\tilde{g}(Z, \theta, \gamma)]$ 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 they produce a value of $E_\pi[\tilde{g}(Z, \theta, \gamma)]$ that is also close to 0 within a specified tolerance. Hence, solutions at infinity do not require a separate treatment in practice.

The condition $E_\pi[\tilde{g}(Z, \theta, \gamma)] = 0$ is the first order condition of a convex optimization problem in γ (this follows from $\partial E_\pi[\tilde{g}(Z, \theta, \gamma)]/\partial\gamma'$ being positive-definite, as shown in 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 $E_\pi[\tilde{g}(Z, \theta, \gamma)]$ will reach its unique global minimum for γ such that $E_\pi[\tilde{g}(Z, \theta, \gamma)] = 0$ and has no other local minima (this can be shown by writing $\partial(g_\gamma' W g_\gamma)/\partial\gamma = 2(\partial g_\gamma'/\partial\gamma) W g_\gamma$, where $g_\gamma \equiv E_\pi[\tilde{g}(Z, \theta, \gamma)]$, and where both $\partial g_\gamma'/\partial\gamma$ and W are positive-definite, so $2(\partial g_\gamma'/\partial\gamma) W g_\gamma = 0$ if and only if $g_\gamma = 0$).

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 $\mu \in \mathcal{P}_{U|Z}$ of the unobservables that satisfy the moment condition. Since we only need to find one suitable μ, it is useful to rank the possible μ's using some convenient objective function and to convert an abstract "existence problem" into a more concrete optimization problem. If there exists no $\mu \in \mathcal{P}_{U|Z}$ that satisfy the moment conditions, the optimization problem will find no solution. If there exists a unique $\mu \in \mathcal{P}_{U|Z}$ that satisfies the moment conditions, the maximization will find it. If there exist more than one $\mu \in \mathcal{P}_{U|Z}$ (or even infinitely many) that satisfy 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

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

To rank distributions in $\mathcal{M}_{\theta,\pi}$, we use entropy, since it has a long history as a 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 that are not possible with other intuitive choices.

Generally, the entropy S of a distribution is defined relative to a reference measure, say $\rho \in \mathcal{P}_{U|Z}$ (which may depend on θ, although this is suppressed in the notation for simplicity), as[11]

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

Among all $\mu \in \mathcal{M}_{\theta,\pi}$, we select the one that maximizes this quantity:[12]

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

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

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

We can set up a Lagrangian for this optimization problem,

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

where $f(u|z) \equiv d\mu(u|z)/d\rho(u|z)$ 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.,

$$
\int d\mu(u|z) = 1 \quad \text{or} \quad \int \bigl(d\mu(u|z)/d\rho(u|z)\bigr)\, d\rho(u|z) = 1
$$

for π-almost every $z \in \mathcal{Z}$). This infinite-dimensional constraint also ensures that the 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 $\delta f(u|z)$:

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

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

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

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

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

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

$$
\int\int g(u, z, \theta) f(u|z)\, d\rho(u|z)\, d\pi(z) = 0,
$$

that is,

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

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 that solves (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 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 that satisfies 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 to maximize the likelihood $\int\int \ln(\frac{d\mu(u|z)}{d\rho(u|z)})\, d\rho(u|z)\, d\pi(z)$ (instead of (8)). As shown in Appendix E of the Supplemental 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 Appendix E of the Supplemental 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 $\int\int (\frac{d\mu(u|z)}{d\rho(u|z)})^2\, d\rho(u|z)\, d\pi(z)$, 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.[13]

2.3. *Estimation Outline*

The simplest way to evaluate the integral (7) that defines 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, for example, 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,

[13]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.

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 equality

$$\begin{aligned}\tilde{g}(z,\theta,\gamma) = &\int g(u,z,\theta)\exp\bigl(\gamma' g(u,z,\theta) - \gamma_0' g(u,z,\theta_0)\bigr)\\ &\times r(u|z,\theta_0,\gamma_0)\,d\rho(u|z;\theta_0)\\ &\Big/\int \exp\bigl(\gamma' g(u,z,\theta) - \gamma_0' g(u,z,\theta_0)\bigr)\\ &\times r(u|z,\theta_0,\gamma_0)\,d\rho(u|z;\theta_0),\end{aligned}$$

where

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

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:

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

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 further consider such issues here.

Averaging over the unobservables then provides us with a conventional moment condition $E[\tilde{g}(Z,\theta,\gamma)] = 0$ that involves only observable variables and

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 (2007)). Empirical likelihood is also known to exhibit desirable optimal power properties under large deviation criteria in the context of point-identified models (Kitamura (2001), Kitamura, Santos, and Shaikh (2012)) and in a large class of set-identified models (Canay (2010)). While statistical optimality criteria point toward one-step methods, GMM offers one convenient computational advantage: Its objective function involves some sample averages that are linear in $\tilde{g}(z, \theta, \gamma)$, 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 $\hat{g}(z, \theta, \gamma)$.

The possibility of set identification (rather than point identification) will require special attention when calculating confidence regions. Appendix F of the Supplemental 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 (S.13) in Appendix C.1 are specific examples of this. In special cases (i.e., when $\text{Closure}\{E_{\mu\times\pi}[g(U, Z, \theta)] : \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}\}$ 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*

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

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[14] (Andrews and Shi (2013), Kim (2008), Menzel (2008), Molinari (2008), Chernozhukov, Lee, and Rosen (2013)), 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), 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 to deal 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 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

[14]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 $\mathcal{T}$).

are combined with the observed data and fed into a conventional generalized method of moments (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 that 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 those 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 that covers 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 that describe 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 exhibits 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 Appendix I of the Supplemental 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 Barwick (2012), 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 that involve unobservables are known to imply moment inequality constraints that can be derived by exploiting linearity or monotonicity (see, among many others, Manski (1995, 2003), Magnac and Maurin (2008), Molinari (2008), Example 1.5 above, and the general approach of Bontemps, Magnac, and Maurin (2012)). 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 Barwick (2012), Canay (2010)). However, in general, obtaining equivalent inequalities is not a trivial problem. Our ex-

plicit 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 contrast, 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 (2013) and Ekeland, Galichon, and Henry (2010). Like the present approach, their method manages to replace an infinite-dimensional nuisance parameter with a finite-dimensional one.[15] 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 method (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 that define 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 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 surprising that Galichon and Henry (2013) only provide very simple simulation examples of latent variable models where the inequalities are not known in advance (with at most two moment conditions and discrete observables). Similarly, Beresteanu, Molchanov, and Molinari (2011) focused on examples where the support function is a maximum

[15]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 $G(u_g, \theta) = \{z : u_g = g(u, z, \theta), u \in \mathcal{U}\}$.

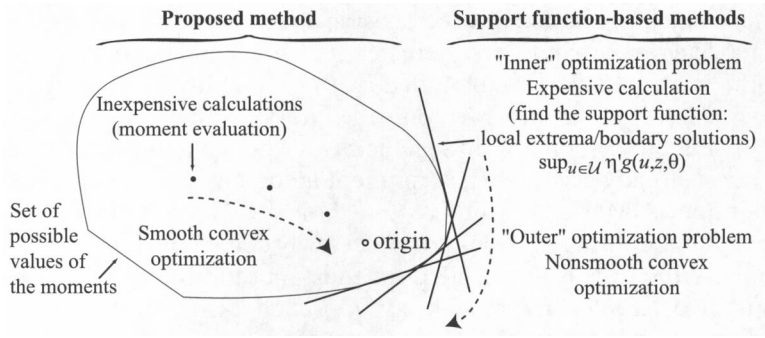

FIGURE 1.—Comparison of the proposed method with methods based on support functions (Galichon and Henry (2013), Ekeland, Galichon, and Henry (2010), Beresteanu, Molchanov, and Molinari (2011)). To reduce the dimensionality of the problem so that it can be pictured, this figure considers a simple case where the observable Z is constant and where there are only two moment conditions.

over a discrete set of at most three 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 to obtaining numerically exact inequalities that define 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.[16] 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 (2013) and Ekeland, Galichon, and Henry (2010) is that their method does not cover unbounded moment

[16]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 to "interior point methods" in convex optimization (Boyd and Vandenberghe (2004, Chapter 11)).

functions[17] (such as the mean of a variable with unbounded support, a fairly common occurrence). Similarly, Beresteanu, Molchanov, and Molinari (2011) also worked with sets that are bounded (with probability 1). 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 that lead 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 $\overline{\mathcal{U}}$ that has a cardinality greater than 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 $E[\mathring{g}(\mathring{U}, Z, \theta)] = 0$, where $\mathring{g}(\mathring{u}, z, \theta) = g(m(\mathring{u}, z, \theta), z, \theta)$ and where $\mathring{U}$ is a random variable that has 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 $g(u, z, \theta)$.

[17]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.

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

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

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

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

This approach can be fully generalized. If

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

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

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

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

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

Inference regarding θ can then be carried out using an objective function where the nuisance parameters ϕ have been "profiled" out (see Appendix F in the Supplemental 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

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

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))

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

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 B.1 in Appendix B of the Supplemental 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 $\tilde{t}$)

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

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 $E[g_j(u, z, \theta, \nu)] = 0$ 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 B.2 of Appendix B of the Supplemental Material.

We now state an infinite-dimensional version of Theorem 2.1 that 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 $\rho(\cdot|\cdot; \theta, \nu^{(J)}) \in P_{\mathcal{U}|\mathcal{Z}}$ be a user-specified conditional measure that satisfies (i) $\operatorname{supp}\rho(\cdot|z; \theta, \nu^{(J)}) = \mathcal{U}$ for each $z \in \mathcal{Z}$ and (ii) that $E_\pi[\ln E_{\rho(\cdot|\cdot;\theta,\nu^{(J)})}[\exp(\sum_{j=1}^J \gamma_j g_j(u, z, \theta, \nu))|Z]]$ 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 that potentially depend 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*

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

and

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

where

$$\tilde{g}^{(J)}\big(Z,\theta,\nu^{(J)},\gamma\big) \equiv \int g^{(J)}\big(u,z,\theta,\nu^{(J)}\big) \times \exp\big(\gamma' g^{(J)}\big(u,z,\theta,\nu^{(J)}\big)\big)\, d\rho\big(u|z;\theta,\nu^{(J)}\big) \Big/ \int \exp\big(\gamma' g^{(J)}\big(u,z,\theta,\nu^{(J)}\big)\big)\, d\rho\big(u|z;\theta,\nu^{(J)}\big),$$

where $\rho(\cdot|\cdot;\theta,\nu^{(J)})$ *is as in Definition* 4.1 *and* $g^{(J)}(u,z,\theta,\nu^{(J)}) = [g_j(u,z,\theta,(\nu^{(J)},0))]_{j=1}^{J}$ *in which* $\nu^{(J)} \in \mathcal{V}^{(J)}$ *denotes the elements of* $\nu \in \mathcal{V}$, *upon which* $g^{(J)}(u,z,\theta,\nu^{(J)})$ *depends. Then* (i) $\Theta_0^{(J+1)} \subseteq \Theta_0^{(J)}$, (ii) $\bigcap_{J\in\mathbb{N}^*}\Theta_0^{(J)} = \Theta_0$, *and* (iii) $d_H(\Theta_0^{(J)},\Theta_0) \to 0$ *as* $J \to \infty$, *where* $d_H(\mathcal{A},\mathcal{B}) \equiv \max\{\sup_{\alpha\in\mathcal{A}}\inf_{\beta\in\mathcal{B}}\|\alpha-\beta\|, \sup_{\beta\in\mathcal{B}}\inf_{\alpha\in\mathcal{A}}\|\alpha-\beta\|\}$ *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 (2009), 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 that yield 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, that is, $u = (v_{t-l}, v_t)$, 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 parameters per moment condition to go to infinity as the 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 samples, 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 Appendix F of the Supplemental 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*

Appendices C.1 and C.2 of the Supplemental 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 Appendix H of the Supplemental 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 that gives rise to the bounds may be associated with implausible patterns of heteroskedasticity in the residuals that can be restricted by adding moment conditions to ensure 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 par-

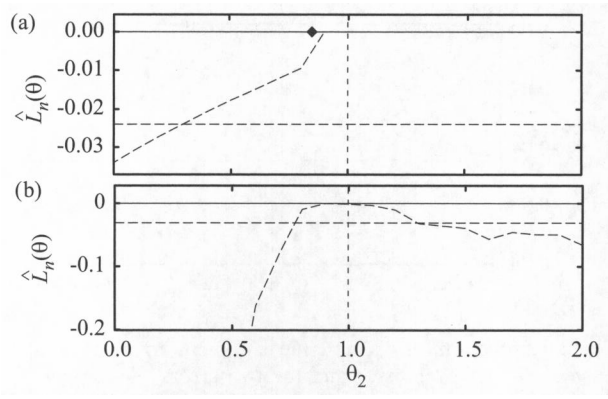

FIGURE 2.—Objective function for a censored regression model. (a) Result obtained with the usual uncorrelatedness and zero mean assumptions on the residuals. The upper diamond marks the well known analytic lower bound for this model. (b) Same exercise while assuming, in addition, that the variance of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

ticularly 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 independent and identically distributed (i.i.d.) 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 (e.g., Klepper and Leamer (1984)) for this model.[18]

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

[18]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.

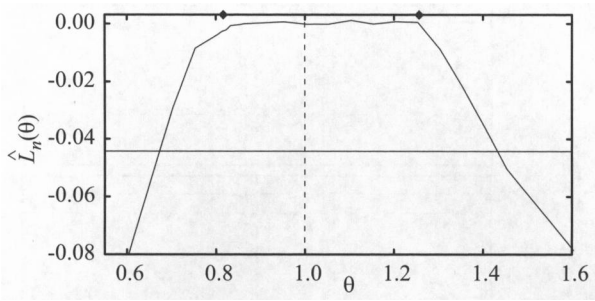

FIGURE 3.—Objective function for a simple measurement error model assuming mutual uncorrelatedness between the true regressor and the two errors. The upper diamonds mark the standard forward and reverse regression bounds for this model. The horizontal line indicates the critical value (at the 95% level)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

point-identified models. We use a sample of 250 i.i.d. 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 cases (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., $E[(X^*)^2(V_1)^2] = E[(X^*)^2]E[(V_1)^2]$). 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 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

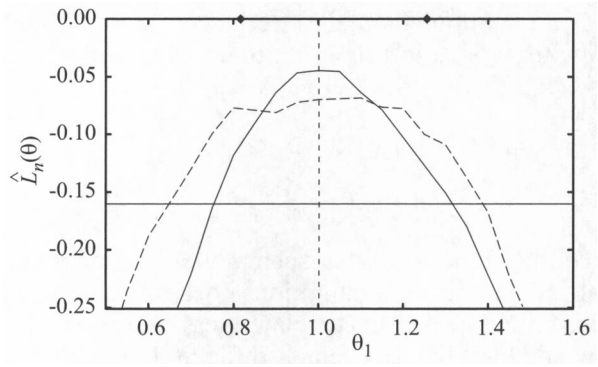

FIGURE 4.—Objective function for the measurement error model under the assumption of mutual independence between the regressors and the two errors. The dashed curve is for the set-identified case with a normally distributed regressor, while the solid curve is for the point-identified case with a uniformly distributed regressor. The horizontal line is 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.

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 that 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 overidentified (all moment conditions need not be satisfied in a given sample). This is not pathological—it is merely a clear indication that the model 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 overidentified.[19]

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.

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

Appendix C.3 of the Supplemental 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 forms

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

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

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 that handles the standard measurement error problem were needed.

6. CONCLUSION

This paper introduces a generalization of GMM to moments that involve 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.

APPENDIX 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 *and* $\tilde{g}$ *as in Theorem* 2.1,

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

and

$$V_\gamma = \int \left(\int \big(g(u,z,\theta) - \tilde{g}(z,\theta,\gamma)\big)\big(g(u,z,\theta) - \tilde{g}(z,\theta,\gamma)\big)' \times \exp\big(\gamma' g(u,z,\theta)\big)\, d\rho(u|z) \Big/ \int \exp\big(\gamma' g(u,z,\theta)\big)\, d\rho(u|z) \right) d\pi(z) \tag{20}$$

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* $E_\pi[\ln E_\rho[\exp(\gamma' g(U,Z,\theta))|Z]]$ *and* $V_\gamma = \partial g_\gamma / \partial \gamma'$.

PROOF: See Appendix B of the Supplemental Material. *Q.E.D.*

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

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

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 that contains 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 $M(\lambda)$ and $\sup \operatorname{supp} F$ would just be shifted by that same constant, since the multiplicative shift in the exponentials cancels in the ratio that defines $M(\lambda)$). The conclusion is equivalent to showing that for $\lambda > 0$ sufficiently large, $M(\lambda)$ 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

$$M(\lambda) = \frac{\int_{x<c} x \exp(\lambda x)\, dF(x) + \int_{x \geq c} x \exp(\lambda x)\, dF(x)}{\int \exp(\lambda x)\, dF(x)} \tag{21}$$

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

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

$$= \frac{-\int_{x<c} |x|\, dF(x) \exp(\lambda c) + c \int_{x\geq c} \exp(\lambda x)\, dF(x)}{\int_{x<c} \exp(\lambda x)\, dF(x) + \int_{x\geq c} \exp(\lambda x)\, dF(x)}$$

$$= \frac{-\int_{x<c} |x|\, dF(x) + c\left[\int_{x\geq c} \exp(\lambda(x-c))\, dF(x)\right]}{\int_{x<c} \exp(\lambda(x-c))\, dF(x) + \left[\int_{x\geq c} \exp(\lambda(x-c))\, dF(x)\right]}.$$

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

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

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 $\int_{x<c} \exp(\lambda(x-c))\, dF(x) \leq \int_{x<c} dF(x) \leq 1$)

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

Hence, we have shown that any b in the interior of $\mathcal{H}$ will eventually be exceeded 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. *Q.E.D.*

PROOF OF THEOREM 2.1: For given $\theta \in \Theta$ and $\pi \in \mathcal{P}_{\mathcal{Z}}$, let

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

be 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

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

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 $0 \in \mathcal{K}_{\theta,\pi}$ if and only if

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 here are justified, and g_γ, V_γ, and V_γ^{-1} exist for all finite values of γ. Then

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

from which we can conclude that $\|g_{\gamma(t)}\|^2 = \|g_0\|^2 \exp(-t)$. 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 $z \in \mathcal{Z} \setminus \mathcal{Z}_+$, 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[20] 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, $\|\gamma\| = r \to \infty$, if and only if 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 that joins g_0 to 0 (see Figure 5(a)), 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 5(b)), the process would stop because $\gamma(t)$ would

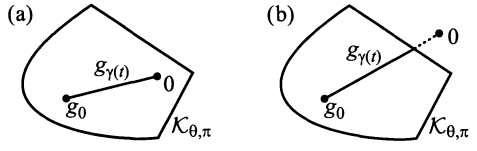

FIGURE 5.—(a) Path of $g_{\gamma(t)}$ when the origin is contained in $\mathcal{K}_{\theta,\pi}$. (b) Path of $g_{\gamma(t)}$ when the origin is not contained in $\mathcal{K}_{\theta,\pi}$.

[20]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 $r_0 \in \mathbb{R}$.

diverge to infinity before g_γ could reach 0. By definition, $g_0 \in \mathcal{K}_{\theta,\pi}$ (because $g_0 = E_{\mu\times\pi}[g(U, Z, \theta)]$ for $\mu = \rho$). Since $\mathcal{K}_{\theta,\pi}$ is closed and convex, the segment that joins g_0 to 0 is entirely contained in $\mathcal{K}_{\theta,\pi}$ if and only if 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 $E_{\mu\times\pi}[g(U, Z, \theta)]$ for $\mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$, the process only fails if $\inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \|E_{\mu\times\pi}[g(U, Z, \theta)]\| \neq 0$. *Q.E.D.*

PROOF OF COROLLARY 2.1: Apply Theorem 2.1 to the moment function $\mathring{g}(u, z, \mathring{\theta}) \equiv g(u, z, \theta) - \phi$ with $\mathring{\theta} \equiv (\theta, \phi)$. For any given θ, the identified set for ϕ gives the range of possible values of $E[g(u, z, \theta)]$. *Q.E.D.*

PROOF OF THEOREM 2.2: Let $\mathcal{K}_\theta = \text{Closure}\{E_{\mu\times\pi_0}[g(U, Z, \theta)] : \mu \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}\}$, the set of all possible values of the moment conditions, and note that by definition, $0 \in \mathcal{K}_\theta$ if and only if $\theta \in \Theta_0$. 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)),

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

where $\bar{t}(\theta, \eta)$ is a scalar function (the so-called support function of the set $\mathcal{K}_\theta$) that we will now determine through $\bar{t}(\theta, \eta) = \sup_{\kappa\in\mathcal{K}_\theta} \eta'\kappa$. Note that we allow $\bar{t}(\theta, \eta)$ 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 $\bar{t}(\theta, \eta)$ 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 that satisfy the inequality.) By Corollary 2.1, $\mathcal{K}_\theta$ is also equal to $\text{Closure}\{E_{\pi_0}[\tilde{g}(Z, \theta, \gamma)] : \gamma \in \mathbb{R}^{d_g}\}$. Therefore, $\bar{t}(\theta, \eta) = \sup_{\gamma\in\mathbb{R}^{d_g}} E_{\pi_0}[\eta'\tilde{g}(Z, \theta, \gamma)]$. Considering one value of z and applying Lemma A.2 for a fixed z with $X = \eta'g(U, z, \theta)$ and F equal to the distribution of $\eta'g(U, z, \theta)$ implies that

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

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, using monotone convergence (as in the proof of Theorem 2.1),

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

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

PROOF OF THEOREM 4.1: By Theorem 2.1, at each finite J, it is clear that

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

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

$$\inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \sup_{j\in\mathbb{N}^*} \left|\int\int g_j(u, z, \theta, \nu)\, d\mu(u|z)\, d\pi(z)\right| \neq 0$$

and there exists a J_0 such that for all $J \ge J_0$

$$\inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \sup_{j\in\{0,\dots,J\}} \left|\int\int g_j(u, z, \theta, \nu)\, d\mu(u|z)\, d\pi(z)\right| \neq 0.$$

If $\theta \in \Theta_0$, then

$$\inf_{\nu\in\mathcal{V}} \inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}} \left|\int\int g_j(u, z, \theta, \nu)\, d\mu(u|z)\, d\pi(z)\right| = 0$$

for all $j \in \mathbb{N}^*$. It follows that (ii) $\bigcap_{J\in\mathbb{N}^*} \Theta_0^{(J)} = \Theta_0$.

Finally, $d_H(\Theta_0, \Theta^{(J+1)}) = d_H(\bar{\Theta}_0, \bar{\Theta}^{(J+1)})$ since closure does not affect the Hausdorff distance. Also, $d_H(\bar{\Theta}_0, \bar{\Theta}^{(J+1)}) = \sup_{\theta\in\bar{\Theta}^{(J+1)}} \inf_{\tilde{\theta}\in\bar{\Theta}_0} \|\theta - \tilde{\theta}\|$ because $\sup_{\theta\in\bar{\Theta}_0} \inf_{\tilde{\theta}\in\bar{\Theta}^{(J+1)}} \|\theta - \tilde{\theta}\| = 0$ since $\bar{\Theta}_0 \subset \bar{\Theta}^{(J+1)}$. Next, $d_H(\bar{\Theta}_0, \bar{\Theta}^{(J+1)}) \le \sup_{\theta\in\bar{\Theta}^{(J)}} \inf_{\tilde{\theta}\in\bar{\Theta}_0} \|\theta - \tilde{\theta}\| = d_H(\bar{\Theta}_0, \bar{\Theta}^{(J)})$ since $\bar{\Theta}^{(J+1)} \subseteq \bar{\Theta}^{(J)}$. Since $d_H(\bar{\Theta}_0, \bar{\Theta}^{(J)})$ forms a nonincreasing sequence and $d_H(\bar{\Theta}_0, \bar{\Theta}^{(J)}) \ge 0$, we have $d_H(\bar{\Theta}_0, \bar{\Theta}^{(J)}) \to c \ge 0$. We now show that c must be 0. We must have $\sup_{\theta\in\bar{\Theta}^{(J)}} \inf_{\tilde{\theta}\in\bar{\Theta}_0} \|\theta - \tilde{\theta}\| \ge c$ 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 $\|\theta_J - \tilde{\theta}_J\| \ge c$ 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 $\theta_\infty \in \bigcap_{J\in\mathbb{N}^*} \bar{\Theta}^{(J)}$, for otherwise eventually θ_∞ would lie at a finite distance from $\bar{\Theta}^{(J)}$ and $\|\theta_{J_j} - \theta_\infty\| \not\to 0$. 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 $\|\theta_\infty - \tilde{\theta}_\infty\| = 0$ and that $c = 0$. Hence $d_H(\bar{\Theta}_0, \bar{\Theta}^{(J)}) \to 0$. *Q.E.D.*

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. [348]

ANDREWS, D. W. K., AND P. J. BARWICK (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826. [346,363]

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. [346,363]

ANDREWS, D. W. K., AND X. SHI (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666. [362]

ANDREWS, D. W. K., AND G. SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [346,363]

BERESTEANU, A., AND F. MOLINARI (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763–814. [346]

BERESTEANU, A., I. MOLCHANOV, AND F. MOLINARI (2011): "Sharp Identification Regions in Models With Convex Moment Predictions," *Econometrica*, 79, 1785–1821. [362-366]

BIERENS, H. J. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458. [367]

BLUNDELL, R., M. BROWNING, AND I. CRAWFORD (2008): "Best Nonparametric Bounds on Demand Responses," *Econometrica*, 76, 1227–1262. [348]

BONTEMPS, C., T. MAGNAC, AND E. MAURIN (2012): "Set Identified Linear Models," *Econometrica*, 80, 1129–1155. [363]

BOYD, S., AND L. VANDENBERGHE (2004): *Convex Optimization*. New York: Cambridge University Press. [365]

BUGNI, F. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78, 735–753. [346,363]

CANAY, I. (2010): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," *Journal of Econometrics*, 156, 408–425. [346,361,363]

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334. [367-369]

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284. [346-348,361,363,370, 373]

CHERNOZHUKOV, V., E. KOCATULUM, AND K. MENZEL (2012): "Inference on Sets in Finance," Working Paper, Department of Economics, Massachusetts Institute of Technology. [346]

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737. [362]

CHIBURIS, R. C. (2008): "Approximately Most Powerful Tests for Moment Inequalities," Working Paper, Princeton University. [346,363]

CRAGG, J. C. (1997): "Using Higher Moments to Estimate the Simple Errors-in-Variables Model," *RAND Journal of Economics*, 28, S71–S91. [372]

CRESSIE, N., AND T. R. C. READ (1984): "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society, Ser. B.*, 46, 440–464. [359]

CSISZAR, I. (1975): "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, 3, 146–158. [358]

——— (1991): "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *The Annals of Statistics*, 19, 2032–2066. [356]

DONALD, S. G., G. IMBENS, AND W. NEWEY (2009): "Choosing Instrumental Variables in Conditional Moment Restriction Models," *Journal of Econometrics*, 152, 28–36. [369]

DUDLEY, R. (2002): *Real Analysis and Probability*. New York: Cambridge University Press. [350]

EKELAND, I., A. GALICHON, AND M. HENRY (2010): "Optimal Transportation and the Falsifiability of Incompletely Specified Economic Models," *Economic Theory*, 42, 355–374. [346,352,362, 364,365]

ENDOU, N., K. NARITA, AND Y. SHIDAMA (2008): "The Lebesgue Monotone Convergence Theorem," *Formalized Mathematics*, 16, 167–175. [379]

GALICHON, A., AND M. HENRY (2013): "Dilation Bootstrap," *Journal of Econometrics*, 177, 109–115. [346,364,365]

GEWEKE, J., AND M. KEANE (2001): "Computationally Intensive Methods for Integration in Econometrics," in *Handbook of Econometrics*, Vol. V. Amsterdam: Elsevier. [360]

GOLAN, A., G. JUDGE, AND D. MILLER (1996): *Maximum Entropy Econometrics: Robust Estimation With Limited Data*. New York: Wiley. [356,363]

GOURIEROUX, C., AND A. MONFORT (1997): *Simulation-Based Econometric Methods*. New York: Oxford University Press. [360]

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700. [355]

HAILE, P. A., AND E. TAMER (2003): "Inference With an Incomplete Model of English Auctions," *Journal of Political Economy*, 111, 1–51. [348]

HAJIVASSILIOU, V. A., AND P. A. RUUD (1994): "Classical Estimation Methods for LDV Models Using Simulation," in *Handbook of Econometrics*, Vol. IV. Amsterdam: Elsevier, 2384–2438. [360]

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054. [345,347]

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. Amsterdam: Elsevier. [369]

HARVEY, A., S. J. KOOPMAN, AND N. SHEPHARD (2004): *State Space and Unobserved Component Models: Theory and Applications*. Cambridge: Cambridge University Press. [369]

HAUSMAN, J., W. NEWEY, AND J. POWELL (1995): "Nonlinear Errors in Variables: Estimation of Some Engel Curves," *Journal of Econometrics*, 65, 205–233. [373]

IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357. [346,356]

KIM, K. (2008): "Set Estimation and Inference With Models Characterized by Conditional Moment Inequalities," Working Paper, University of Minnesota. [362]

KIRKPATRICK, S., C. D. GELATT, AND M. P. VECCHI (1983): "Optimization by Simulated Annealing," *Science*, 220, 671–680. [360,365]

KITAMURA, Y. (2001): "Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 69, 1661–1672. [361]

KITAMURA, Y., AND M. STUTZER (1997): "An Information-Theoretic Alternative to Generalized Method of Moment Estimation," *Econometrica*, 65, 861–874. [346]

KITAMURA, Y., A. SANTOS, AND A. M. SHAIKH (2012): "On the Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 80, 413–423. [361]

KLEPPER, S., AND E. E. LEAMER (1984): "Consistent Sets of Estimates for Regressions With Errors in All Variables," *Econometrica*, 52, 163–183. [371]

KULLBACK, S. (1959): *Information Theory and Statistics*. New York: Wiley. [356]

MAGNAC, T., AND E. MAURIN (2008): "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data," *Review of Economic Studies*, 75, 835–864. [363]

MANSKI, C. (1995): *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press. [363]

——— (2003): *Partial Identification of Probability Distributions*. New York: Springer. [363]

MANSKI, C. F., AND E. TAMER (2002): "Inference on Regressions With Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546. [347]

MCFADDEN, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995–1026. [354,359,360,362]

MCFADDEN, D. L. (2005): "Revealed Stochastic Preference: A Synthesis," *Economic Theory*, 26, 245–264. [348]

MENZEL, K. (2008): "Estimation and Inference With Many Moment Inequalities," Working Paper, Massachusetts Institute of Technology. [362]

MOLINARI, F. (2008): "Partial Identification of Probability Distributions With Misclassified Data," *Journal of Econometrics*, 144, 81–117. [362,363]

NEWEY, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627. [363]

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engel and D. L. McFadden. Amsterdam: Elsevier. [347]

NEWEY, W., AND R. J. SMITH (2004): "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255. [346,361]

OWEN, A. B. (1988): "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249. [346]

——— (1990): "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120. [346]

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057. [354,359,360,362]

PAKES, A., J. PORTER, K. HO, AND J. ISHII (2005): "Moment Inequalities and Their Application," Working Paper, Harvard University. [348]

PONOMAREVA, M., AND E. TAMER (2011): "Misspecification in Moment Inequality Models: Back to Moment Equalities?" *Econometrics Journal*, 14, 186–203. [347]

QIN, J., AND J. LAWLESS (1994): "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22, 300–325. [346]

ROCKAFELLAR, R. T. (1970): *Convex Analysis*. Princeton: Princeton University Press. [380]

ROEHRIG, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447. [352]

ROMANO, J. P., AND A. M. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211. [346,363]

ROSEN, A. M. (2008): "Confidence Sets for Partially Identified Parameters That Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107–117. [346,363]

SCHENNACH, S. M. (2007): "Point Estimation With Exponentially Tilted Empirical Likelihood," *The Annals of Statistics*, 35, 634–672. [346,361]

——— (2014): "Supplement to 'Entropic Latent Variable Integration via Simulation'," *Econometrica Supplemental Material*, 82, http://www.econometricsociety.org/ecta/supmat/9748_miscellaneous.pdf; http://www.econometricsociety.org/ecta/supmat/9748_data_and_programs.zip. [347]

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. [374]

SHEN, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591. [363]

SHEN, X., J. SHI, AND W. H. WONG (1999): "Random Sieve Likelihood and General Regression Models," *Journal of the American Statistical Association*, 94, 835–846. [363]

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. [356]

STINCHCOMBE, M. B., AND H. WHITE (1998): "Consistent Specification Testing With Nuisance Parameters Present Only Under the Alternative," *Econometric Theory*, 14, 295–325. [367]

VARIAN, H. R. (1982): "The Nonparametric Approach to Demand Analysis," *Econometrica*, 50, 945–973. [348]

ZELLNER, A. (1997): "The Bayesian Method of Moments (BMOM)," *Advances in Econometrics*, 12, 85–105. [356]

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

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