

# EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity

Ivan A. Canay<sup>\*†</sup>

*Department of Economics, Northwestern University*

This Draft: October 2008

## Abstract

This paper addresses the issue of optimal inference for parameters that are partially identified in models with moment inequalities. There currently exists a variety of inferential methods for use in this setting. However, the question of choosing optimally among contending procedures is unresolved. In this paper, I first consider a canonical large deviations criterion for optimality and show that inference based on the empirical likelihood ratio statistic is optimal. This finding is a direct analog to that in Kitamura (2001) for moment *equality* models. Second, I introduce a new empirical likelihood bootstrap that provides a valid resampling method for moment inequality models and overcomes the implementation challenges that arise as a result of non-pivotal limit distributions. Lastly, I analyze the finite sample properties of the proposed framework using Monte Carlo simulations. The simulation results are encouraging.

**Keywords:** empirical likelihood, partial identification, large deviations, empirical likelihood bootstrap, optimal inference.

**JEL Classification:** C12, C14, C21.

---

<sup>\*</sup>Email: [iacanay@northwestern.edu](mailto:iacanay@northwestern.edu) - Website: <http://faculty.wcas.northwestern.edu/~iac879/>. First Draft: April 2006.

<sup>†</sup>I am deeply grateful to Jack Porter and Bruce Hansen for thoughtful discussions and constant encouragement. I also want to acknowledge helpful conversations with Dennis Kristensen, Ken West, Steven Durlauf, Don Andrews and Elie Tamer as well as the good feedback provided by seminar participants at the 2007 Summer Meeting of the Econometric Society, UW-Madison, UCL, Northwestern, Brown, Boston University, Columbia, Harvard, Duke, Chicago GSB, Penn State and Michigan. Comments are welcome.

# 1 Introduction

Recently, there have been many papers proposing methodologies for estimation and inference in models where the parameter of interest is not uniquely defined by the economic model and the distribution of the observed data (see, among others, Chernozhukov, Hong, and Tamer 2007, Pakes, Porter, Ho, and Ishii 2005, Romano and Shaikh 2008, Romano and Shaikh 2006, Imbens and Manski 2004, Rosen 2008, Beresteanu and Molinari 2008). When this situation arises the model is said to be partially identified. Given this expanding literature on various inferential methods, it is natural to wonder which method is optimal. This paper addresses the question of optimal inference and contains the following contributions. First, I consider a canonical large deviations criterion for optimality and show that inference based on the empirical likelihood ratio (ELR) statistic is optimal. Second, I introduce a simple and natural modification of the empirical likelihood bootstrap introduced by Brown and Newey (2002) that provides a valid bootstrap method for moment inequality models. This modified empirical likelihood bootstrap is important to overcome the implementation challenges associated with non-pivotal limit distributions in partially identified models. Third, I conduct a Monte Carlo experiment which suggests a finite sample performance advantage to this bootstrap procedure as well. These results firmly ground empirical likelihood as an attractive method for inference in moment inequality models.

The problem of optimal inference can be interpreted as a problem of optimal choice of a criterion function. Partially identified models are usually represented via a population objective function  $Q(\theta, P_0)$  which does not have a unique minimizer, so that

$$\Theta_0(P_0) = \arg \min_{\theta \in \Theta} Q(\theta, P_0) \tag{1.1}$$

represents a set containing all the values of  $\theta$  consistent with the economic model and the distribution  $P_0$ . The primary goal is to use a sample analog  $\widehat{Q}$  of  $Q(\theta, P_0)$  to construct confidence regions that cover each of the elements of  $\Theta_0(P_0)$  with a given probability. Most of these models involve a moment inequality condition of the form  $\mathbb{E}[m(z, \theta)] \geq 0$  in which case  $\Theta_0(P_0)$  is the set of all  $\theta$  that satisfy the moment condition.<sup>1</sup> In such cases, there are many different choices of  $Q(\theta, P_0)$  that have  $\Theta_0(P_0)$  as the minimizer set and each choice could lead to different sample analogs and thus different confidence sets. The question of interest is whether there is an optimal criterion function  $Q^*(\theta, P_0)$ , where optimal means that inference based on  $\widehat{Q}^*$  is more *precise* than inference based on any other sample criterion function. This paper contributes to the growing literature on inference in partially identified models by giving an answer to this question. I introduce empirical likelihood (EL) as a new procedure for partially identified models and show that inference based on the empirical likelihood ratio (ELR) statistic is optimal in a large deviations sense.

---

<sup>1</sup>Some binary choice models as well as situations in which the lack of identification is related to the failure of some rank condition (see Santos 2006, Chernozhukov, Hong, and Tamer 2007) cannot be represented as a moment inequality.

The method of empirical likelihood is known to have several optimality properties for models with *equality* moment restrictions. In terms of point estimation, the EL estimator is semiparametrically efficient (i.e., attains the semiparametric efficiency bound derived by Chamberlain 1987). In addition, this estimator often exhibits desirable properties in terms of higher order comparisons (see Newey and Smith 2004). Regarding inference, the empirical likelihood ratio test has some desirable features as well. DiCiccio, Hall, and Romano (1991) proved that the ELR test admits Bartlett correction, which gives the same accuracy rate as the parametric case. Kitamura (2001) used the so-called Generalized Neyman-Pearson approach to show that for testing moment restrictions the ELR test is uniformly most powerful in an asymptotic large deviations sense. Additional results on minimax risk optimality are presented by Kitamura and Otsu (2005). This is just a sample of the large list of papers that show some sort of optimality for EL. The interested reader is referred to Kitamura (2006) and Owen (2001) for additional discussion. Also, Moon and Schorfheide (2008) consider equalities as well as inequalities in their estimation procedure. However, their model is assumed to be point identified by the set of equality restrictions.

The search for an optimal inferential procedure in partially identified models involves a number of additional complications that are not found in the point identified case. The fact that  $\Theta_0(P_0)$  is no longer a singleton complicates the use of local asymptotic optimality notions since standard expansion tools are not as obviously available.<sup>2</sup> Another optimality notion that has been widely applied in point identified models is the large deviations approach. This approach has the virtue of translating naturally to the partially identified setting and is the criterion I pursue here.

The theory of large deviations deals with the behavior of estimators in a fixed neighborhood of the true value. Suppose that there is a statistic  $T_n$  that converges in probability to  $T$  and let  $A$  denote a set such that the closure of  $A$  does not contain  $T$ . For each  $n$ ,  $\Pr(T_n \in A) \rightarrow 0$ . In typical cases,  $\Pr(T_n \in A) \rightarrow 0$  at an exponential rate, i.e. there exists a constant  $0 < \eta < \infty$  such that,

$$n^{-1} \log \Pr(T_n \in A) \rightarrow -\eta. \quad (1.2)$$

Notice the contrast with conventional local asymptotic theory where the focus is on the behavior of  $T_n$  in a shrinking neighborhood of the true parameter value,  $T$ . Here the neighborhood  $A$  is fixed. For example, let  $X_1, \dots, X_n$  be an *i.i.d.* sample of standard normal random variables and consider the empirical mean  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ . Since  $\bar{X}_n$  is also normal with zero-mean and variance  $1/n$ , for any  $\delta > 0$ ,

$$P(|\bar{X}_n| \geq \delta) = 1 - (\sqrt{2\pi})^{-1} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-x^2/2} dx \rightarrow 0 \quad \implies \quad \frac{1}{n} \log P(|\bar{X}_n| \geq \delta) \rightarrow -\frac{\delta^2}{2} \quad (1.3)$$

---

<sup>2</sup>This is in fact related to the lack of an asymptotic distribution for sets. The methodology used by Beresteanu and Molinari (2008) is a promising direction for this type of analysis. Also, Andrews and Soares (2007) use a local asymptotic analysis to compare the power properties of different critical values for a given criterion function.

Equation (1.3) is an illustration of a large deviations statement: the typical value of  $\bar{X}_n$  is of order  $1/\sqrt{n}$ , but with small probability (of the order  $e^{-n\delta^2/2}$ ),  $\bar{X}_n$  takes values outside a fixed bound. The study of the large deviations behavior of the type I and type II error probabilities associated with any given test procedure gives insight on the good performance of the test: the higher the rate of decrease of these errors, the better the inference. Thus, while standard definitions of efficiency (e.g., Pitman relative efficiency) make the testing problem harder by considering alternative hypothesis that get closer to the null hypothesis as the sample size increases, the efficiency criteria based on large deviations make the problem harder by letting the type I and type II error probabilities go to zero asymptotically. Precise definitions and statements are postponed to section 3, where I show that the ELR test achieves the fastest rates of decrease.

The second main contribution of this paper is related to the actual implementation of the new optimal procedure. I show that under the null hypothesis the ELR statistic converges to a well defined asymptotic distribution. However, whether this limit distribution is actually useful to calculate critical values depends on the case under consideration. The number of binding constraints - the components of the vector  $m(z, \theta)$  with zero expectation - depends crucially on  $\theta$  and this causes the asymptotic distribution to be non-pivotal. The non-pivotalness is not a barrier in some cases and then one can compute valid critical values using a simple and straightforward approximation. In more complicated setups though, these approximations could be really slack (see Wolak 1991, Guggenberger, Hahn, and Kim 2008) so that using a resampling technique could be desirable. One alternative that many authors have adopted in these types of models is the use of subsampling for the construction of critical values. Subsampling would in fact be valid to approximate the limit distribution of the ELR statistic. Yet, the contribution in this paper lies in a different alternative. I first show that the empirical likelihood bootstrap proposed by Brown and Newey (2002) is not asymptotically valid when applied to moment inequality models.<sup>3</sup> Then I propose a slight modification of that bootstrap, along the lines of the modified parametric bootstrap in Andrews (2000), that does work asymptotically. The modification involves changing the set of inequalities  $\mathbb{E}[m(z, \theta)] \geq 0$  by  $\mathbb{E}[m(z, \theta)] \geq \varrho_n$ , where  $\varrho_n$  is a positive sequence that goes to zero asymptotically.<sup>4</sup>

Before proceeding any further, I mention the recent literature that has introduced different techniques to deal with partially identified models and is closely related to the tools presented here. Horowitz and Manski (1998, 2000) and later Imbens and Manski (2004) developed methods for estimation and inference for the case where the identification region is defined by lower and upper bounds that can be estimated from the data. This situation usually arises under the presence of missing data (see Example 1), which is a common phenomenon in many statistical applications. Manski and Tamer (2002) consider regression models with interval data and show that this also results in parameters that are only partially identified. For an

---

<sup>3</sup>The EL bootstrap proposed by Brown and Newey (2002) was developed for models comprised of moment *equality* restrictions.

<sup>4</sup>This idea is related to the independently derived work by Andrews and Soares (2007) and Bugni (2007).

excellent exposition of these data issues see Manski (2003). Going beyond these particular cases, Chernozhukov, Hong, and Tamer (2004, 2007), Andrews, Berry, and Jia (2004) and Pakes, Porter, Ho, and Ishii (2005) were the first to extend the methodologies to more general setups. Andrews, Berry, and Jia (2004) analyze economic models of entry (see Example 2); Pakes, Porter, Ho, and Ishii (2005) consider structural models in which the equilibrium conditions impose inequality moment conditions; and Chernozhukov, Hong, and Tamer (2007) define the identified set as the solution of the minimization of a criterion function and provide several results on estimation and inference on both  $\theta_0$  and  $\Theta_0(P_0)$ . Romano and Shaikh (2008) and Romano and Shaikh (2006) carry out a further analysis of the validity of subsampling and present conditions under which the confidence regions cover the parameter of interest uniformly. For additional results on uniform coverage see Soares (2006) and the recent papers by Andrews and Guggenberger (2008) and Andrews and Soares (2007). Rosen (2008) presents a connection between moment inequality models and the literature on one-sided hypothesis testing. As it will be noted in the next section, his Gaussian quasi-likelihood ratio (QLR) statistic is closely related to the empirical likelihood ratio statistic proposed here. Using a different line of analysis Beresteanu and Molinari (2008) propose an inference procedure for partially identified models that can be written as a transformation of an expectation of a set valued random variable. Galichon and Henry (2006b) and Galichon and Henry (2006a) address the choice of the criterion function  $Q(\theta, P_0)$  suggesting the use of a Kolmogorov-Smirnov-type statistic and present a bootstrap algorithm to make feasible inference in their setup. Finally, Bugni (2007) presents a bootstrap procedure for partially identified models.<sup>5</sup> All the papers already mentioned present theoretical contributions and, in some cases, small scale applications. The works by Ciliberto and Tamer (2006), Ho (2005) and Ishii (2005) are some of the papers that apply the tools mentioned above.

It is worth noticing that both the empirical likelihood approach for partially identified models and the modified empirical likelihood bootstrap introduced in this paper involve changes that preserve the computational simplicity associated with the method of empirical likelihood for point identified models. Also, the bootstrap confidence regions proposed in this paper are uniform in the sense of Andrews and Soares (2007). Finally, Monte Carlo simulations show a good performance of the new EL bootstrap in finite samples.

The remainder of the paper is organized as follows. Section 2 presents the basic notation and the two canonical examples used throughout the paper. Section 3 develops the empirical likelihood approach for unconditional moment inequalities and shows that inference based on the empirical likelihood ratio statistic is large deviations optimal. There are two main results in the section. The first one characterizes the asymptotic distribution of the ELR statistic. The second one deals with the so-called Generalized Neyman-Pearson optimality and extends the results in Kitamura (2001) to the moment inequality case. Once the optimal properties

---

<sup>5</sup>The bootstrap procedure proposed in this paper and in Bugni (2007) were independently derived. Both procedures are closely related to the GMS approach introduced, also independently, by Andrews and Soares (2007).

are defined, section 4 introduces the new empirical likelihood bootstrap for moment inequality models. Section 5 studies the finite sample behavior of the method proposed in this paper via Monte Carlo simulations. Finally, section 6 concludes.

## 2 Setup and Notation

The point of departure of this paper is a statistical model that imposes an inequality moment condition. To write this formally, let  $\{z_i\}_{i=1}^n$  be an i.i.d. sample generated from some distribution  $P_0 \in \mathcal{M}$  with support on  $\mathcal{Z} \subseteq \mathbb{R}^d$ ,  $\theta_0$  be the parameter of interest that lies in  $\Theta \subseteq \mathbb{R}^k$  and  $m(z, \theta)$  be a  $q \times 1$  known function. The economic model imposes,

$$\mathbb{E}_{P_0}[m(z, \theta_0)] \equiv \int_{\mathcal{Z}} m(z, \theta_0) dP_0 \geq 0, \quad (2.1)$$

and so the true value of the parameter might not be uniquely identified by the distribution of the observed data  $P_0$  and the economic model. Under the moment condition (2.1), the set

$$\Theta_0(P_0) \equiv \{\theta \in \Theta \subseteq \mathbb{R}^k : \mathbb{E}_{P_0}[m(z, \theta)] \geq 0\} \quad (2.2)$$

denotes the so-called *identified set* while any  $\theta \in \Theta_0(P_0)$  has been termed an *identifiable parameter* by Romano and Shaikh (2008).

**Example 1** *Missing data - Manski (1989)*. Let  $\{(x_i; w_i) : i = 1, \dots, n\}$  be a random sample with support  $[0, 1] \times \{0, 1\}$ , where  $w_i$  is always observed and  $x_i$  is only observed if  $w_i = 1$ . The propensity score  $p = \mathbb{E}(w_i) \in (0, 1)$  is unknown to the researcher, but is consistently estimated by its sample analog. The goal is to make inferences on  $\theta_0 = \mathbb{E}(x)$ , which is alternatively described as  $\theta_0 = p \cdot \mu_1 + (1 - p) \cdot \mu_0$ , where  $\mu_1 = \mathbb{E}(x|w = 1)$  and  $\mu_0 = \mathbb{E}(x|w = 0)$ . Notice that while  $\mu_1$  is identified by the observed data,  $\mu_0$  is only known to be in the interval  $[0, 1]$ . Therefore, this model yields the following two moment inequalities,

$$\begin{aligned} \mathbb{E}[m_1(x, w, \theta_0)] &= \mathbb{E}[\theta_0 - xw] \geq 0 \\ \mathbb{E}[m_2(x, w, \theta_0)] &= \mathbb{E}[1 - w + xw - \theta_0] \geq 0 \end{aligned}$$

and the identified set is simply  $\Theta_0(P_0) = [\theta_L, \theta_H]$  where  $\theta_L = p \cdot \mu_1$  and  $\theta_H = p \cdot \mu_1 + (1 - p)$ .

**Example 2** *Entry model - Tamer (2003)*. Suppose that firm  $j \in \{1, 2\}$  decides whether to enter ( $z_{j,m} = 1$ ) a market  $m \in \{1, \dots, M\}$  or not ( $z_{j,m} = 0$ ) based on the profit function  $\pi_{j,m} = (\varepsilon_{j,m} - \theta_j z_{-j,m}) 1\{z_{j,m} = 1\}$ , where  $\varepsilon_{j,m}$  is firm's  $j$  benefit of entry in market  $m$  and  $z_{-j,m}$  denotes the decision of the other firm. Let  $\varepsilon_{j,m} \sim U(0, 1)$  and  $\theta_0 = (\theta_1, \theta_2) \in (0, 1)^2$ . There are four possible outcomes in this game: (i) if  $\varepsilon_{j,m} > \theta_j$  for  $j = \{1, 2\}$  the unique Nash equilibrium (NE) is  $(z_{1,m}, z_{2,m}) = (1, 1)$ ; (ii) if  $\varepsilon_{1,m} > \theta_1$  and  $\varepsilon_{2,m} < \theta_2$  the unique NE is  $(z_{1,m}, z_{2,m}) = (1, 0)$ ; (iii) if  $\varepsilon_{1,m} < \theta_1$  and  $\varepsilon_{2,m} > \theta_2$  the unique NE is  $(z_{1,m}, z_{2,m}) = (0, 1)$  and;

(iv) if  $\varepsilon_{j,m} < \theta_j$  for  $j = \{1, 2\}$  there are multiple equilibria as both  $(z_{1,m}, z_{2,m}) = (1, 0)$  and  $(z_{1,m}, z_{2,m}) = (0, 1)$  are NE. Without imposing additional assumptions this model implies

$$\begin{aligned} P(z_{1,m} = 1, z_{2,m} = 1) &= (1 - \theta_1)(1 - \theta_2) \\ \theta_2(1 - \theta_1) \leq P(z_{1,m} = 1, z_{2,m} = 0) &\leq \theta_2, \end{aligned}$$

so that there are two moment inequalities and one moment equality,

$$\begin{aligned} \mathbb{E}[m_1(z_m, \theta_0)] &= \mathbb{E}[z_{1,m}z_{2,m} - (1 - \theta_1)(1 - \theta_2)] = 0 \\ \mathbb{E}[m_2(z_m, \theta_0)] &= \mathbb{E}[z_{1,m}(1 - z_{2,m}) - \theta_2(1 - \theta_1)] \geq 0 \\ \mathbb{E}[m_3(z_m, \theta_0)] &= \mathbb{E}[\theta_2 - z_{1,m}(1 - z_{2,m})] \geq 0. \end{aligned}$$

The identified set  $\Theta_0(P_0)$  is given by,

$$\Theta_0(P_0) = \left\{ \theta \in \Theta : \begin{array}{l} \mathbb{E}[z_{1,m}(1 - z_{2,m})] \leq \theta_2 \leq \frac{\mathbb{E}[z_{1,m}(1 - z_{2,m})]}{\mathbb{E}[z_{1,m}(1 - z_{2,m})] + \mathbb{E}[z_{1,m}z_{2,m}]} \\ \theta_1 = 1 - \mathbb{E}[z_{1,m}z_{2,m}]/(1 - \theta_2) \end{array} \right\}.$$

With the set of moment inequalities at hand, the goal is to construct confidence regions  $\mathcal{C}_n$  that contain the parameter of interest with at least some prespecified probability. Depending on the case, interest might center on the single element  $\theta_0$  or on  $\Theta_0(P_0)$ . That is, in partially identified models there is a distinction between inference on the identified set or on individual elements of that set (see, for example, Imbens and Manski 2004, Chernozhukov, Hong, and Tamer 2007, Romano and Shaikh 2006). This paper focuses on inference about an identifiable parameter  $\theta$ . This focus corresponds to interest in a particular “true value” of the parameter, which often has a particular economic interpretation. Another distinction of importance is whether the coverage of the confidence region is only valid for a fixed probability distribution  $P_0$  or if the coverage is uniform over a large class of probability functions  $\mathcal{M}$ . Using the terminology in Romano and Shaikh (2008), confidence regions in the former case are called *pointwise consistent in levels*<sup>6</sup> while confidence regions that satisfy the latter requirement are referred as *uniformly consistent in levels*. This paper focuses on confidence regions  $\mathcal{C}_n$  that are pointwise consistent in levels, i.e.,

$$\inf_{\theta \in \Theta_0(P_0)} \lim_{n \rightarrow \infty} P_0(\theta \in \mathcal{C}_n) \geq 1 - \alpha. \quad (2.3)$$

However, the EL confidence regions proposed in this paper are also uniformly valid in the sense of Andrews and Soares (2007).<sup>7</sup>

Construction of a confidence region  $\mathcal{C}_n$  with the required coverage level typically proceeds as follows. One uses a sample criterion function  $\widehat{Q}(\theta)$  associated with a population criterion

<sup>6</sup>In this case, for every sample size  $n$  there is a probability distribution for which the coverage of the confidence region is not close to the nominal level.

<sup>7</sup>This result is included in a supplementary appendix available upon request.

function  $Q(\theta, P_0)$  and exploits the duality between hypothesis tests and confidence sets. The confidence region arises by inverting the test of each of the individual null hypotheses  $H_0^\theta : \theta \in \Theta_0(P_0)$  so that  $\mathcal{C}_n \equiv \{\theta \in \Theta : \widehat{Q}(\theta) \leq c\}$  for a given a cut-off value  $c$ . In this paper I show that the use of the empirical likelihood ratio statistic  $\mathcal{ELR}_n(\theta)$  results in optimal inference when testing the hypothesis  $H_0^\theta : \mathbb{E}_{P_0}[m(z_i, \theta)] \geq 0$ , which is equivalent to  $H_0^\theta : \theta \in \Theta_0(P_0)$ . I also show that  $\mathcal{ELR}_n(\theta)$  has a well defined asymptotic distribution that allows one to compute an asymptotic critical value  $c_{1-\alpha}^\theta$  for a pre-specified asymptotic size  $\alpha$ . Thus, for each  $\theta \in \Theta_0(P_0)$  there is a  $c_{1-\alpha}^\theta$  such that

$$\lim_{n \rightarrow \infty} P_0(\mathcal{ELR}_n(\theta) > c_{1-\alpha}^\theta) \leq \alpha, \quad (2.4)$$

meaning that

$$\mathcal{C}_n \equiv \{\theta \in \Theta : \mathcal{ELR}_n(\theta) \leq c_{1-\alpha}^\theta\} \quad (2.5)$$

satisfies (2.3). As will be clear later on, in some cases the asymptotic distribution might not be that useful to compute  $c_{1-\alpha}^\theta$ . For such cases I introduce a modified empirical likelihood bootstrap that consistently estimates  $c_{1-\alpha}^\theta$ .<sup>8</sup>

### 3 Optimal Inference in Partially Identified Models

It is useful to think of the search for an optimal inferential method in terms of choice of criterion function.<sup>9</sup> This is so because a model that imposes  $\mathbb{E}_{P_0}[m(z, \theta)] \geq 0$  has many different criterion functions  $Q(\theta, P_0)$  that have  $\Theta_0(P_0)$  as the minimizer set. This fact translates into consequences for the inferential procedure given that a standard method of confidence region construction is to invert a test based on a sample analog of the criterion function. Therefore, each criterion function results in a different test statistic, a different decision rule and so a different confidence region. This section shows that appropriate choice of  $Q(\theta, P_0)$  leads to a large deviations optimal inferential procedure. Consider the next example for an illustration.

**Example 3 (Missing Data cont.)** Define,  $m(z, \theta)' = [m_1(z, \theta) \ m_2(z, \theta)]$  where  $m_1(z, \theta) = \theta - xw$  and  $m_2(z, \theta) = 1 - w + xw - \theta$ . The identified set is  $\Theta_0(P_0) = [\theta_L, \theta_H]$ . Below I give four different criterion functions that provide an equivalent representation of the identified set via  $\Theta_0(P_0) = \arg \min_{\theta \in \Theta} Q_k(\theta, P_0)$ , for  $k = \{1, \dots, 4\}$ .

1.  $Q_1(\theta, P_0) \equiv (\mathbb{E}_{P_0}[m(z, \theta)])' W(\theta) (\mathbb{E}_{P_0}[m(z, \theta)])_-$ . Here  $(x)_- = \min\{x, 0\}$  and  $W(\theta)$  is a positive definite weighting matrix. This criterion function for the choice  $W(\theta) = I_{2 \times 2}$  has been used by Chernozhukov, Hong, and Tamer (2007) and Romano and Shaikh (2008).
2.  $Q_2(\theta, P_0) \equiv \min_{t \geq 0} (\mathbb{E}_{P_0}[m(z, \theta)] - t)' V(\theta)^{-1} (\mathbb{E}_{P_0}[m(z, \theta)] - t)$ , where  $V(\theta) = \text{var}\{m(z, \theta)\}$ . This Gaussian quasi-likelihood ratio criterion function has been used by

<sup>8</sup>The ELR statistic also satisfies the conditions required to use subsampling as in Chernozhukov, Hong, and Tamer (2007) and Romano and Shaikh (2008).

<sup>9</sup>However, the optimality criterion consider below is general and do not restrict us to inferential methods based on sample analogs of criterion functions.



Rosen (2008).

3.  $Q_3(\theta, P_0) \equiv \sup_{A \in S} (P_0(A) - v_\theta(\Gamma_\theta(A)))^2$ . This is the Kolmogorov-Smirnov statistic proposed by Galichon and Henry (2006b), where  $v_\theta$  is the probability measure of the unobservables,  $\Gamma_\theta$  is a correspondence from observables to unobservables and  $S$  is a determining class of sets.
4.  $Q_4(\theta, P_0) \equiv \max_{\lambda \leq 0} \mathbb{E}_{P_0}[\log(1 + \lambda' m(z, \theta))]$ . This criterion function defines the identified set (Lemma B.3 in Appendix B) and is the one associated with empirical likelihood.

Example 3 raises the following question: is there a criterion function that is *better* than the rest? In this paper I define *better* in terms of asymptotic power since more powerful tests reject more false hypotheses and generally lead to smaller confidence regions. In particular, I study the large deviations behavior of statistics that serve to test the null hypotheses  $H_0^\theta : \theta \in \Theta_0(P_0)$ . I use a large deviations optimality criteria that is defined in terms of asymptotic power and show that tests based on the empirical likelihood ratio statistic are optimal according to this criteria. The result then gives a justification for the use of  $\mathcal{ELR}_n(\theta)$ , the statistic associated with the criterion function  $Q_4(\theta, P_0)$  in example 3.

One notion of optimality for confidence regions is to focus on the optimality of the corresponding tests that are inverted to form the confidence region. To assess the relative or absolute performance of test procedures there exist several optimality criteria which, in order to make informative comparisons, usually consider problems that become harder as more observations become available. A line of attack that is applicable to a wide range of cases is based on the theory of large deviations and has been used since the papers by Bahadur (1960), Chernoff (1952) and Hoeffding (1965), among others. Thus, test procedures are compared through their power functions and the various methods of comparison differ in the manner in which type I and type II error probabilities vary with the sample size, and also with respect to the manner in which the alternatives under consideration are required to behave. Letting  $\alpha_n$  and  $\beta_n$  denote the type I and type II error probabilities of a test, each performance criteria entails particular specifications regarding: (i) whether  $\alpha_n$  goes to zero or not, (ii) whether  $\beta_n$  goes to zero or not, and (iii) whether the alternative hypotheses are fixed or get closer to the null with the sample size.<sup>10</sup>

In this paper I focus on large deviations cases where both types of error probabilities decrease to zero as the sample size increases and the set of alternatives is held fixed. I follow Kitamura (2001) and use the so-called Generalized Neyman-Pearson approach, an extension of the idea introduced by Hoeffding (1965) for multinomial models. To put it in simple terms, consider competing tests that satisfy

$$\limsup_{n \rightarrow \infty} n^{-1} \log \alpha_n \leq -\eta,$$

---

<sup>10</sup>For a review of comparison of test procedures see Serfling (1980, Ch. 10) and van der Vaart (1998, Ch. 14)

for a given  $\eta > 0$ . Among such tests, a test is optimal if it minimizes

$$\limsup_{n \rightarrow \infty} n^{-1} \log \beta_n$$

uniformly over all distributions in a given class with the same support. In section 3.3 I show that the empirical likelihood-based test of the unconditional moment restrictions (2.1) is optimal in the Generalized Neyman-Pearson sense.<sup>11</sup> In anticipation of such result and for ease of exposition, I first introduce empirical likelihood for moment inequalities.

### 3.1 The ELR Statistic for Moment Inequalities

EL is a data-driven nonparametric method of estimation and inference for moment restriction models, which does not require weight matrix estimation like GMM<sup>12</sup> and is invariant to nonsingular linear transformations of the moment conditions. It was introduced by Owen(1988, 1990, 1991) and later studied in depth by Qin and Lawless (1994), Imbens, Spady, and Johnson (1998), Kitamura (2001) and Newey and Smith (2004), among others.

The standard EL for moment *equalities* maximizes the nonparametric likelihood over distributions with an atom of probability on each  $z_i$  that impose the moment condition. In this paper I use this idea, but impose the moment *inequality* (2.1) into the optimization problem. The (restricted) empirical log-likelihood problem is,

$$l_{EL}^r(\theta) \equiv \max_{p_1, \dots, p_n} \left\{ \sum_{i=1}^n \log(p_i) \left| p_i \geq 0; \sum_{i=1}^n p_i = 1; \sum_{i=1}^n p_i m(z_i, \theta) \geq 0 \right. \right\} \quad (3.1)$$

where  $p_i$  denotes the probability mass placed at  $z_i$  by a discrete distribution with support  $\{z_1, \dots, z_n\}$ . Note that this differs from the usual EL for moment equalities only in the last restriction where  $\sum_{i=1}^n p_i m(z_i, \theta)$  is now required to satisfy an inequality. The ELR statistic arises by computing the difference between the restricted and unrestricted log-likelihood,

$$\mathcal{ELR}_n(\theta) \equiv 2\{l_{EL}^{ur}(\theta) - l_{EL}^r(\theta)\}, \quad (3.2)$$

where the unrestricted empirical log-likelihood problem  $l_{EL}^{ur}(\theta)$  is similar to  $l_{EL}^r(\theta)$  except that the moment restriction  $\sum_{i=1}^n p_i m(z_i, \theta) \geq 0$  is not imposed. Consequently, large values of this statistic suggest that such a restriction is not supported by the sample.

A nice feature of EL is that imposing moment inequalities preserves the simplicity of the optimization problem from the moment equality case. The only difference lies in the behavior

---

<sup>11</sup>The ELR statistic is also Chernoff-minimax optimal as in Puhalskii and Spokoiny (1998) and Kitamura and Otsu (2005). This result is included in a supplementary appendix available upon request.

<sup>12</sup>This is true even computationally when ones uses algorithms with numerical derivatives.

of the Lagrange multipliers. To see this, note that (3.1) is solved by maximizing the Lagrangian

$$\mathbb{L}(\theta, \{p_i\}_{i=1}^n, \lambda, \varkappa) \equiv \sum_{i=1}^n \log(p_i) + \varkappa \left(1 - \sum_{i=1}^n p_i\right) - n\lambda' \sum_{i=1}^n p_i m(z_i, \theta) \quad (3.3)$$

where  $\varkappa$  is the Lagrange multiplier for the second constraint and  $\lambda \leq 0$  is a  $q \times 1$  vector of multipliers for moment inequality constraints. Solving this problem results in

$$\hat{p}_i = \frac{1}{n(1 + \hat{\lambda}' m(z_i, \theta))}, \quad (3.4)$$

which looks identical to the standard EL solution. The difference lies in  $\hat{\lambda}$ , which now solves the following three first order conditions<sup>13</sup>

$$\sum_{i=1}^n \frac{m(z_i, \theta)}{n(1 + \hat{\lambda}' m(z_i, \theta))} \geq 0; \quad \hat{\lambda} \leq 0; \quad \hat{\lambda}' \sum_{i=1}^n \frac{m(z_i, \theta)}{n(1 + \hat{\lambda}' m(z_i, \theta))} = 0. \quad (3.5)$$

Since  $\hat{p}_i$  has a closed form solution, I can write a profiled likelihood and define  $\hat{\lambda}$  accordingly. Plugging (3.4) into (3.3) results in

$$l_{EL}^r(\theta) \equiv \mathbb{L}(\theta, \{\hat{p}_i\}_{i=1}^n, \lambda, \hat{\varkappa}) = \min_{\lambda \leq 0} \left\{ -n \log(n) - \sum_{i=1}^n \log(1 + \lambda' m(z_i, \theta)) \right\}. \quad (3.6)$$

Going back to the unrestricted problem  $l_{EL}^{ur}(\theta)$ , note that the solution to that case is  $\tilde{p}_i = 1/n$  so that  $l_{EL}^{ur}(\theta) = -n \log(n)$  and then,

$$\mathcal{ELR}_n(\theta) \equiv 2\{l_{EL}^{ur}(\theta) - l_{EL}^r(\theta)\} = \max_{\lambda \leq 0} 2 \sum_{i=1}^n \log(1 + \lambda' m(z_i, \theta)). \quad (3.7)$$

In order to derive the asymptotic distribution of  $\mathcal{ELR}_n(\theta)$  it is convenient to write an alternative parametrization such that the derivative with respect to  $\lambda$  is zero at the optimum. One way to do this is by introducing an additional multiplier for the restriction  $\lambda \leq 0$ . Denote this multiplier by  $\tau \geq 0$  and define an alternative empirical likelihood ratio statistic as:

$$\begin{aligned} \mathcal{ELR}_n(\theta) &= \min_{\tau \geq 0} \max_{\lambda(\theta, \tau) \in \mathbb{R}^q} \left\{ 2 \sum_{i=1}^n \log(1 + \lambda(\theta, \tau)' m(z_i, \theta)) - 2\lambda(\theta, \tau)' n\tau \right\} \\ &= \min_{\tau \geq 0} \max_{\lambda(\theta, \tau) \in \mathbb{R}^q} R_n(\theta, \lambda, \tau). \end{aligned} \quad (3.8)$$

Now the derivative with respect to  $\lambda$  is zero,  $\frac{\partial}{\partial \lambda} R_n(\theta, \lambda, \tau) = 0$ . In addition, the derivative with respect to  $\tau$  is non-negative,  $\frac{\partial}{\partial \tau} R_n(\theta, \lambda, \tau) \geq 0$ , and this in turn imposes  $\lambda \leq 0$  and the third moment condition in (3.1). Note also that  $\hat{\tau} = \sum_{i=1}^n \hat{p}_i m(z_i, \theta) \geq 0$ .

Therefore, a model represented through moment inequalities affects the empirical likelihood

---

<sup>13</sup> $\hat{\lambda}$  is a shorthand for  $\hat{\lambda}(\theta)$  or  $\hat{\lambda}(\theta, \tau)$  depending on the context.

ratio statistic only via the Lagrange multiplier  $\lambda$  which is now required to be non-positive. This difference is important for two reasons. First, relative to the standard case, the computational difficulty is only trivially affected. Second, the restriction on  $\lambda$  affects the limit distribution of the statistic to a great extent and results in a non-pivotal asymptotic distribution as the next section shows.

### 3.2 Asymptotic Distribution of the ELR Statistic

Before deriving the asymptotic distribution of the ELR statistic I introduce some additional notation of the model. Let  $m(z_i, \theta)$  be  $q \times 1$  with the following partition:  $m(z_i, \theta)' = [m_b(z_i, \theta)' m_s(z_i, \theta)']$  where  $m_b(z_i, \theta)$  is a  $b(\theta) \times 1$  vector of moments with zero mean,  $\mathbb{E}_{P_0}[m_b(z_i, \theta)] = 0$ , and  $m_s(z_i, \theta)$  is a  $s(\theta) \times 1$  vector of moments with positive mean,  $\mathbb{E}_{P_0}[m_s(z_i, \theta)] > 0$ . The asymptotic behavior of the test statistic will follow from the next standard assumption.

**Assumption 3.1** (i)  $\{z_i\}_{i=1}^n$  is a random sample where  $z_i \in \mathcal{Z} \subseteq \mathbb{R}^d$ . (ii)  $\Theta \subseteq \mathbb{R}^k$  is compact and  $m : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^q$  is known. (iii)  $\sup_{\theta \in \Theta_0(P_0)} \mathbb{E}_{P_0}[||m(z_i, \theta)||^r] < \infty$  for  $r \geq 3$ . (iv) For each  $\theta \in \Theta_0(P_0)$ ,  $\Sigma_b(\theta) = \mathbb{E}_{P_0}[m_b(z_i, \theta)m_b(z_i, \theta)']$  is positive definite.

Assumption 3.1(iii) implies that for all  $\theta \in \Theta_0(P_0)$ ,

$$\sqrt{n}\{\bar{m}_n(\theta) - \mathbb{E}_{P_0}[m(z_i, \theta)]\} \rightsquigarrow N(0, V(\theta))$$

where  $\bar{m}_n(\theta) = \frac{1}{n} \sum_{i=1}^n m(z_i, \theta) \xrightarrow{p} \mathbb{E}_{P_0}[m(z_i, \theta)]$ ,  $V(\theta) = \text{var}_{P_0}\{m(z_i, \theta)\}$  and  $\rightsquigarrow$  denotes weak convergence. In particular, for  $m_b(z_i, \theta)$  the same assumption implies  $\sqrt{n}\bar{m}_{b,n}(\theta) \rightsquigarrow N(0, \Sigma_b(\theta))$ .

The following theorem provides a limit distribution for  $\mathcal{ELR}_n(\theta)$  under the null hypothesis.

**Theorem 3.1** Under Assumption 3.1, for all  $\theta \in \Theta_0(P_0)$  the empirical likelihood ratio statistic defined in (3.8) satisfies:

$$\lim_{n \rightarrow \infty} P_0(\mathcal{ELR}_n(\theta) \geq c) = \sum_{j=0}^{b(\theta)} \varpi(b(\theta), b(\theta) - j, \Sigma_b(\theta)) \Pr(\chi_j^2 \geq c) \quad (3.9)$$

where  $\varpi(b(\theta), b(\theta) - j, \Sigma_b(\theta))$  is the weight function defined by Wolak (1987) and Kudo (1963).

**Proof.** See Appendix A. ■

Theorem 3.1 shows that the limit distribution of  $\mathcal{ELR}_n(\theta)$  is a chi-bar-square distribution, which I denote by  $\bar{\chi}_{b(\theta)}^2(\Sigma_b(\theta))$ . This distribution is non-pivotal since both  $b(\theta)$  and  $\varpi(b(\theta), b(\theta) - j, \Sigma_b(\theta))$  depend on  $\theta$ . The set of  $b(\theta)$  binding constraints has a significant and discontinuous effect on the shape of the distribution. In the extreme case where  $\theta$  is such that  $b(\theta) = 0$  - no constraint is binding - the resulting distribution is degenerate at zero since  $\mathcal{ELR}_n(\theta) = 0$  wp $\rightarrow 1$ . The weights  $\varpi(b(\theta), b(\theta) - j, \Sigma_b(\theta))$  are usually called level probabilities

of a chi-bar-square distribution. Aside from the cases where  $b(\theta) \leq 4$ , there are no closed-form expressions for these weights.<sup>14</sup>

Recently, Rosen (2008) derived a QLR statistic for the model in (2.1) and showed that its limit distribution is the chi-bar-square distribution in (3.9). To overcome the problem caused by the non-pivotalness of the statistic Rosen uses two conservative approximations to the chi-bar-square that were proposed by Wolak(1987, 1991). These approximations require imposition of an upper bound  $b^*$  on the number of binding moments  $b(\theta)$ , i.e.,  $\sup_{\theta \in \Theta_0(P_0)} b(\theta) \leq b^*$  - see equation (5.2) in section 5. If  $b^*$  is a good approximation to the supremum and the number of moments is not big, these approximations can work very well. The simulations in section 5 illustrate this feature. However, if the moments are highly non-linear and the number of binding moments is large, these approximations could be rather slack as mentioned by Wolak (1991) and Guggenberger, Hahn, and Kim (2008).

Theorem 3.1 shows that the ELR statistic is equivalent to the QLR statistic proposed by Rosen (2008) up to first order. However, there are two clear advantages in using the methodology of the present paper. First, the analysis of the next subsection shows that  $\mathcal{ELR}_n(\theta)$  outperforms the QLR statistic in terms of asymptotic power and so it yields smaller confidence regions for  $\theta_0$ . Second, the computation of  $\mathcal{ELR}_n(\theta)$  does not require an estimate of  $V(\theta)$  as it is required for the QLR statistic. This feature reflects the internal Studentization property of empirical likelihood that often improves finite sample properties of the tests. In fact, avoiding estimation of  $V(\theta)^{-1}$  is more important than usual in the present setup. For example, in the entry game of section 5 this matrix is singular.<sup>15</sup>

### 3.3 Large Deviations Optimality of the ELR Statistic

I now consider the problem of optimal inference in models that impose the moment inequalities (2.1) and show that the ELR statistic just described yields optimal inference from a large deviations point of view. Recall that by Assumption 3.1,  $\{z_i\}_{i=1}^n$  is an i.i.d. sample generated from some distribution  $P_0 \in \mathcal{M}$  with support on  $\mathcal{Z} \subseteq \mathbb{R}^d$ , where  $\mathcal{M}$  is the space of probability measures on the Borel  $\sigma$ -field  $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ . The null hypothesis for each  $\theta$  is given by

$$H_0^\theta : \mathbb{E}_{P_0}[m(z, \theta)] \equiv \int_{\mathcal{Z}} m(z, \theta) dP_0 \geq 0, \quad (3.10)$$

and from (3.7) the ELR statistic to test this null is,

$$\mathcal{ELR}_n(\theta) \equiv \max_{\lambda(\theta) \leq 0} 2 \sum_{i=1}^n \log(1 + \lambda(\theta)' m(z_i, \theta)). \quad (3.11)$$

---

<sup>14</sup>These issues and much more have been studied in depth in the literature on one-sided hypothesis testing, e.g. Wolak (1987), Silvapulle and Sen (2004) and Gouriéroux, Holly, and Monfort (1982)

<sup>15</sup>Also, inverting  $V(\theta)$  in the missing data example might be problematic when subsampling is used to compute the critical values. When the sample size is small many of the subsamples are such that  $V(\theta)$  is not invertible.

Moreover, for each  $\theta \in \Theta$  let

$$\mathcal{P}_0(\theta) \equiv \{P \in \mathcal{M} : \mathbb{E}_P[m(z_i, \theta)] \geq 0\} \quad (3.12)$$

denote the subset of probability measures that satisfy the moment inequality restriction. With this notation  $H_0^\theta$  can be rewritten as  $H_0^\theta : P_0 \in \mathcal{P}_0(\theta)$ .

The idea of this section works as follows. A test for  $H_0^\theta$  is a map from data into a decision. I denote this mapping by  $r_n^\theta : \mathcal{Z} \rightarrow \{0, 1\}$ , where  $r_n^\theta = 0$  ( $r_n^\theta = 1$ ) means acceptance (rejection) of the null. Characterizing data by the empirical measure  $\hat{P}_n$  it induces, a test maps  $\hat{P}_n$  into accept/reject. This is equivalent to partitioning the space of probability measures into acceptance or rejection regions. Slightly more generally, a test induces a partition  $\Omega_n^\theta \equiv (\Omega_{0,n}^\theta, \Omega_{1,n}^\theta)$  of  $\mathcal{M}$  such that  $r_n^\theta = 1(\hat{P}_n \in \Omega_{1,n}^\theta)$ .

Let  $\bar{r}_n^\theta$  denote the test based on the ELR statistic,

$$\bar{r}_n^\theta \equiv 1 \left( \frac{\mathcal{ELR}_n(\theta)}{2n} \geq \eta^\theta \right). \quad (3.13)$$

The good behavior of tests based on  $\mathcal{ELR}_n(\theta)$  is due to the fact that for each  $\theta \in \Theta$  this statistic is the dual of the following problem,

$$\inf_{P \in \mathcal{P}_0(\theta)} I(\hat{P}_n || P), \quad (3.14)$$

where

$$I(Q || P) \equiv \begin{cases} \int \log(dQ/dP)dQ & \text{if } Q \text{ is absolutely continuous with respect to } P \\ \infty & \text{otherwise} \end{cases} \quad (3.15)$$

denotes the relative entropy (or Kullback-Leibler divergence measure) for probability measures  $P$  and  $Q$ .<sup>16</sup> Intuitively, EL is picking the discrete measure in the set  $\mathcal{P}_0(\theta)$  that is closest to the empirical measure  $\hat{P}_n$ , where closest is defined in terms of the relative entropy. The test  $\bar{r}_n^\theta$  depends on the data solely through the empirical measure and induces the following partition of  $\mathcal{M}$ ,

$$\Lambda_0^\theta \equiv \left\{ Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_0(\theta)} I(Q || P) < \eta^\theta \right\}, \quad \Lambda_1^\theta \equiv \left\{ Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_0(\theta)} I(Q || P) \geq \eta^\theta \right\}. \quad (3.16)$$

Therefore, the ELR test has the following three equivalent representations,

$$\bar{r}_n^\theta \equiv 1 \left( \frac{\mathcal{ELR}_n(\theta)}{2n} \geq \eta^\theta \right) = 1 \left( \inf_{P \in \mathcal{P}_0(\theta)} I(\hat{P}_n || P) \geq \eta^\theta \right) = 1(\hat{P}_n \in \Lambda_1^\theta). \quad (3.17)$$

The main result of this section follows from the two assumptions below, where  $P^n$  denotes

---

<sup>16</sup> $Q$  is absolutely continuous with respect to  $P$  if for each  $\varepsilon > 0$  there exists  $\delta > 0$  such that for each  $B \in \mathcal{B}(\mathcal{Z})$ ,  $P(B) < \delta$  implies  $Q(B) < \varepsilon$ . In this case,  $P$  is said to dominate  $Q$  (sometimes written as  $Q \ll P$ ).

the  $n$ -fold product measure  $\otimes_{i=1}^n P$  of a measure  $P$ .

**Assumption 3.2**  $\mathcal{Z}$  and  $\Theta$  are compact metric spaces.

**Assumption 3.3**  $\mathbb{E}_{P_0}[m(z_i, \theta)]$  is continuous in  $\theta \in \Theta$ .

**Theorem 3.2** Suppose Assumptions 3.1-3.3 hold. For each  $\theta \in \Theta$  and any  $\eta^\theta > 0$ , let  $\Lambda^\theta = (\Lambda_0^\theta, \Lambda_1^\theta)$  be defined as in (3.16). Then,

(I) For all  $P \in \mathcal{P}_0(\theta)$ ,

$$\limsup_{n \rightarrow \infty} n^{-1} \log P^n(\widehat{P}_n \in \Lambda_1^\theta) \leq -\eta^\theta.$$

(II) If a test  $\Omega_n^\theta = (\Omega_{n,0}^\theta, \Omega_{n,1}^\theta)$  satisfies for all  $P \in \mathcal{P}_0(\theta)$

$$\limsup_{n \rightarrow \infty} n^{-1} \log P^n(\widehat{P}_n \in (\Omega_{n,1}^\theta)^\delta) \leq -\eta^\theta, \quad (3.18)$$

for some  $\delta > 0$ , then

$$\limsup_{n \rightarrow \infty} n^{-1} \log P_1^n(\widehat{P}_n \in \Omega_{n,0}^\theta) \geq \limsup_{n \rightarrow \infty} n^{-1} \log P_1^n(\widehat{P}_n \in \Lambda_0^\theta)$$

and

$$\liminf_{n \rightarrow \infty} \frac{P_1^n(\widehat{P}_n \in \Omega_{0,n}^\theta)}{P_1^n(\widehat{P}_n \in \Lambda_0^\theta)} \geq 1 \quad (3.19)$$

for all  $P_1 \in \mathcal{M}$ , where  $(\Omega_{n,1}^\theta)^\delta = \cup_{Q \in \Omega_{n,1}^\theta} B(Q, \delta)$  is an open  $\delta$ -blow up of the set and  $B(Q, \delta)$  is an open ball of radius  $\delta$  around  $Q$ .

**Proof.** See appendix A. ■

**Remark 3.1** As in Kitamura (2001, Theorem 2), theorem 3.2 uses  $\delta$ -smoothing. To get a similar result without the need of smoothing, the alternative test has to be regular (see Kitamura (2001, Corollary 1)). A test  $\Omega_n^\theta$  is regular if

$$\lim_{\delta \rightarrow 0} \sup_{P \in \mathcal{P}_0(\theta)} \limsup_{n \rightarrow \infty} n^{-1} \log P^n(\widehat{P}_n \in (\Omega_{n,1}^\theta)^\delta) = \sup_{P \in \mathcal{P}_0(\theta)} \limsup_{n \rightarrow \infty} n^{-1} \log P^n(\widehat{P}_n \in \Omega_{n,1}^\theta).$$

**Remark 3.2** Note that, as opposed to Theorem 2 in Kitamura (2001), theorem 3.2 does not require  $m(z, \theta)$  to be continuous in  $\theta$ , so it allows  $m(z, \theta)$  to be an indicator function or a simulation-based function.

**Remark 3.3** The notion of optimality in Theorem 3.2 employs a “fixed” (i.e., not data dependent) critical value that depends on  $\theta$  and so is not necessarily conservative. However, in many situations computing such a critical value might not be feasible without using a data-dependent rule.

Theorem 3.2 says that there is no test for the null (3.10) satisfying the rate restriction (3.18) that performs better in term of asymptotic power than EL. In fact, since part (II) of the theorem holds for all  $P_1 \in \mathcal{M}$ , this result is *uniform* in  $\mathcal{M}$  for each  $\theta \in \Theta$ . This finding is not surprising once we understand the importance of the connection between the ELR test and the relative entropy. In order to explain this connection, let  $\{Q_n\}_{n \geq 1}$  be a sequence of probability measures on the Borel  $\sigma$ -field of a Hausdorff topological space  $S$ . In addition, for any given set  $G$  let  $G^\circ$  and  $\bar{G}$  denote the interior and closure, respectively. The large deviation principle (LDP) is defined as follows.

**Definition 3.1 (Large Deviation Principle)** *Let  $I : S \rightarrow [0, \infty]$  be a function such that  $I^{-1}([0, a]) \subset S$  is compact for each  $a > 0$ . The sequence of probability measures  $\{Q_n\}_{n \geq 1}$  is said to obey the LDP with a rate function  $I$  if for any set  $G \subset S$ ,*

$$-\inf_{y \in G^\circ} I(y) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log Q_n(G^\circ) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log Q_n(\bar{G}) \leq -\inf_{y \in \bar{G}} I(y). \quad (3.20)$$

Thus, the LDP characterizes the limiting behavior of a family of probability measures in terms of a rate function  $I(\cdot)$ . For example, the rate function that controls the probability measure of the sample mean in equation (1.3) is  $I(y) = y^2/2$ ,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P_n(\bar{X}_n \in [\delta, \infty)) \leq -\inf_{y \in [\delta, \infty)} \frac{y^2}{2} = -\frac{\delta^2}{2}.$$

As shown by Cramér, the LDP holds in general for sample means of i.i.d. random variables. Since the empirical measure  $\hat{P}_n$  can be viewed as a mean (of Dirac measures), it is a potential candidate for a large-deviation theorem. The first version of such a theorem was proved by Sanov and plays a vital role in the optimality properties of the ELR test so it is stated formally below.

**Theorem 3.3 (Sanov)** *Let  $\mathcal{M}(\Psi)$  denote the space of probability measures on a Polish space  $\Psi$  (i.e., complete separable metric space) equipped with the Lévy metric<sup>17</sup> and take  $P_0 \in \mathcal{M}(\Psi)$ . Then for any set  $G \in \mathcal{M}(\Psi)$ ,*

$$-\inf_{v \in G^\circ} I(v||P_0) \leq \liminf_{n \rightarrow \infty} n^{-1} \log P_0^n(\hat{P}_n \in G^\circ) \leq \limsup_{n \rightarrow \infty} n^{-1} \log P_0^n(\hat{P}_n \in \bar{G}) \leq -\inf_{v \in \bar{G}} I(v||P_0).$$

Therefore, the fundamental reason why tests based on  $\mathcal{ELR}_n(\theta)$  are optimal is the existing connection between  $\mathcal{ELR}_n(\theta)$  and the Kullback-Leibler divergence number. Intuitively, the ELR test uses the minimum Kullback-Leibler divergence between the empirical measure  $\hat{P}_n$  and the set  $\mathcal{P}_0(\theta)$  as a statistical criterion. By Sanov's Theorem, the sequence of empirical measures  $\hat{P}_n$  satisfies the LDP with rate function  $I(Q||P)$ , and then  $I(Q||P)$  controls the limit behavior of the probability that  $\hat{P}_n$  falls into the set  $\mathcal{P}_0(\theta)$ . This is why  $\mathcal{ELR}_n(\theta)$  leads to optimal large deviations procedures.

<sup>17</sup>The Lévy metric is compatible with the weak topology, Dembo and Zeitouni (1998, Theorem D.8).



### 3.3.1 Implications for the Construction of Confidence Regions

The optimality criterion just presented suggests the use of the empirical likelihood ratio statistic for testing the null  $H_0^\theta : \mathbb{E}_{P_0}[m(z, \theta)] \geq 0$  at each  $\theta \in \Theta$ . I now explore how this result translates into good performance of the corresponding confidence regions.

Classical considerations for choosing among confidence regions include accuracy, unbiasedness, equivariance and combinations of these (see Lehmann and Romano 2005). Accuracy seems quite natural: given a pair of confidence procedures with the same probability of covering the correct value, the procedure with smaller chance of covering incorrect values is preferable.<sup>18</sup> Theorem 3.2 has a clear large deviations implication on the accuracy of the empirical likelihood confidence region  $\mathcal{C}_n$ : the good power properties of  $\mathcal{ELR}_n(\theta)$  should be reflected in smaller confidence regions as a more powerful test rejects wrong values of  $\theta$  more often. To be precise, let  $\nu(\cdot)$  denote a measure on  $\Theta$  with respect to the size of  $\Theta_0(P_0)$ ,

$$\nu(\tilde{\mathcal{C}}_n) = \int_{\Theta} 1(\theta \in \tilde{\mathcal{C}}_n \setminus \Theta_0(P_0)) d\nu(\theta)$$

for any  $\tilde{\mathcal{C}}_n \subseteq \Theta$ , where  $\tilde{\mathcal{C}}_n \setminus \Theta_0(P_0) \equiv \tilde{\mathcal{C}}_n \cap \Theta_0^c(P_0)$ . Notice that whenever  $\tilde{\mathcal{C}}_n \subseteq \Theta_0(P_0)$ ,  $\nu(\tilde{\mathcal{C}}_n) = 0$ , so  $\nu(\cdot)$  is a measure of the volume of a confidence region  $\tilde{\mathcal{C}}_n$  that is independent of the size of  $\Theta_0(P_0)$ . In this context, the false coverage of  $\tilde{\mathcal{C}}_n$  can be related to its volume through a version of the so-called Ghosh-Pratt identity [Ghosh (1961), Pratt (1961)]

$$\begin{aligned} \mathbb{E}_{P_0}(\nu(\tilde{\mathcal{C}}_n)) &= \int_{\mathcal{Z}} \int_{\Theta} 1(\theta \in \tilde{\mathcal{C}}_n / \Theta_0(P_0)) d\nu(\theta) dP_0 \\ &= \int_{\Theta} \int_{\mathcal{Z}} 1(\tilde{\mathcal{C}}_n / \Theta_0(P_0) \ni \theta) dP_0 d\nu(\theta) \\ &= \int_{\Theta} P_0(\theta \in \tilde{\mathcal{C}}_n / \Theta_0(P_0)) d\nu(\theta) \\ &= \int_{\Theta_0^c} P_0(\theta \in \tilde{\mathcal{C}}_n) d\nu(\theta), \end{aligned} \tag{3.21}$$

where the last step follows from  $P_0(\theta \in \tilde{\mathcal{C}}_n / \Theta_0(P_0)) = 0$  for all  $\theta \in \Theta_0(P_0)$ . Therefore, the expected volume under  $P_0$  of the confidence region  $\tilde{\mathcal{C}}_n$  equals a sum of the type II error probabilities at  $P_0$  of the associated test of the null hypothesis that  $\theta$  is an identifiable parameter. From the previous section we know that the test based on the empirical likelihood ratio is large deviations most powerful, with a corresponding confidence region given by  $\mathcal{C}_n = \{\theta \in \Theta : \hat{P}_n \in \Lambda_0^\theta\}$  where  $\Lambda_0^\theta = \{Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_0(\theta)} I(Q||P) < \eta^\theta\}$ . Substituting this

---

<sup>18</sup>Accuracy has been widely used as an optimality criterion for confidence regions (see, for example, Cohen and Strawderman 1973, Hooper 1982, Brown, Casella, and Hwang 1995, Evans, Hansen, and Stark 2005). This criterion might not be good in some particular cases though since, as Lehmann and Romano (2005) and Madansky (1962) pointed out, small confidence regions are desirable when they cover the true parameter value but not necessarily otherwise.

into (3.21),

$$\mathbb{E}_{P_0}(\nu(\mathcal{C}_n)) = \int_{\Theta_0^c} P_0(\theta \in \mathcal{C}_n) d\nu(\theta) = \int_{\Theta_0^c} P_0(\widehat{P}_n \in \Lambda_0^\theta) d\nu(\theta). \quad (3.22)$$

Therefore, since for all  $\theta \in \Theta_0^c$  it is the case that  $P_0 \in \mathcal{P}_0^c(\theta) \equiv \mathcal{P}_1(\theta)$  and  $\Lambda_0^\theta$  is uniformly most powerful over  $\mathcal{P}_1(\theta)$  for all  $\theta \in \Theta_0^c$ , the confidence region  $\mathcal{C}_n$  is asymptotically most accurate at  $P_0$  in the following large deviations sense: for any other confidence region  $\widetilde{\mathcal{C}}_n = \{\theta \in \Theta : \widehat{P}_n \in \Omega_{0,n}^\theta\}$  associated with a test  $\Omega_n^\theta$  satisfying the conditions in Theorem 3.2,

$$\limsup_{n \rightarrow \infty} \mathbb{E}_{P_0}(\nu(\widetilde{\mathcal{C}}_n)) \geq \limsup_{n \rightarrow \infty} \mathbb{E}_{P_0}(\nu(\mathcal{C}_n)), \quad (3.23)$$

meaning that the rate at which the expected volume of the EL confidence region decreases is at least as fast as the rate of any other arbitrary confidence region.

## 4 Implementation: A New Empirical Likelihood Bootstrap

This section concerns implementation issues for the ELR statistic. From (2.5) it is clear that once a critical value  $c_{1-\alpha}^\theta$  is available, the construction of the confidence region only involves the evaluation of  $\mathcal{ELR}_n(\theta)$  which is straightforward. The analysis in section 3 shows that  $\mathcal{ELR}_n(\theta)$  has a well defined asymptotic distribution that can be used to compute critical values in some simple cases. More generally, computation of the critical values is infeasible, leading to the use of simulation approximations or resampling techniques. In the present EL setting the empirical likelihood bootstrap introduced by Brown and Newey (2002) for moment *equality* models seems to be an appealing alternative. Since when inequality constraints are present such approach does not produce a consistent approximation to the limit distribution of  $\mathcal{ELR}_n(\theta)$ , as the next section illustrates, I show that a slight modification of this bootstrap is first order valid. The modification is simple enough to preserve the computational tricks that typically make EL straightforward to use.

### 4.1 A Modified Empirical Likelihood Bootstrap

#### 4.1.1 EL Bootstrap Invalidity: A Canonical Example

To motivate the modified empirical likelihood bootstrap, I use a simple example from Romano and Shaikh (2008) to show that the standard empirical likelihood bootstrap does not work. The argument follows the one used by Andrews (2000) to show that the standard i.i.d. bootstrap is inconsistent when the parameter is on the boundary of the parameter space. The example also illustrates intuitively why the new bootstrap does work. The general case is developed following this illustrative example.

Suppose the economic model imposes  $\mathbb{E}_{P_0}(X) \geq \theta$  where  $X_i \sim P_0 = N(0, 1)$ . Without loss of generality and to make the example as simple as possible, I use the criterion function  $Q(\theta, P_0) = (\mathbb{E}_{P_0}(X) - \theta)_-^2$  where  $(a)_- = \min\{a, 0\}$  instead of the one related to the ELR

statistic. The identified set here is  $\Theta_0(P) = [\theta_l, 0]$  where  $\theta_l$  is some lower bound of the parameter space and the sample version of the criterion function satisfies,

$$n\widehat{Q}_n(\theta) = (\sqrt{n}(\overline{X}_n - \theta))_-^2 \rightsquigarrow \begin{cases} 0 & \text{if } \theta < 0 \\ \{\mathbb{Z}\}_-^2 & \text{if } \theta = 0 \end{cases}$$

where  $\mathbb{Z}$  is a standard normal random variable.

The EL bootstrap works as follows. Denote by  $\tilde{P}_n = (\tilde{p}_1, \dots, \tilde{p}_n)$  the EL probabilities that solve

$$l_{EL}^r = \sup_{p_1, \dots, p_n} \left\{ \sum_{i=1}^n \log(p_i) \mid p_i > 0; \sum_{i=1}^n p_i = 1; \sum_{i=1}^n p_i(X_i - \theta) \geq 0 \right\}$$

and let  $\tilde{\mu}_n^\theta \equiv \sum \tilde{p}_i(X_i - \theta) = \max\{\overline{X}_n - \theta, 0\}$ . The bootstrap samples  $\{X_i^*\}_{i=1}^n$  are i.i.d. according to  $\tilde{P}_n = (\tilde{p}_1, \dots, \tilde{p}_n)$  and the bootstrap criterion function is just,

$$n\tilde{Q}_n^*(\theta) = (\sqrt{n}(\overline{X}_n^* - \theta))_-^2.$$

Following Andrews (2000, page 401) let  $B_c \equiv \{\omega : \limsup_{n \rightarrow \infty} \sqrt{n}\overline{X}_n > c\}$  for  $0 < c < \infty$  and note that by the law of iterated logarithm  $P_0(B_c) = 1$ . For  $\omega \in B_c$  consider a subsequence  $\{n_k : k \geq 1\}$  of  $\{n : n \geq 1\}$  such that  $n_k^{1/2}\overline{X}_{n_k}(\omega) \geq c$  for all  $k$ . Then, for such a subsequence,

$$\begin{aligned} n_k\tilde{Q}_{n_k}^*(0) &= (\sqrt{n_k}(\overline{X}_{n_k}^* - \tilde{\mu}_{n_k}^0) + \sqrt{n_k}\tilde{\mu}_{n_k}^0)_-^2 \\ &= (\sqrt{n_k}(\overline{X}_{n_k}^* - \tilde{\mu}_{n_k}^0) + \max\{\sqrt{n_k}\overline{X}_{n_k}, 0\})_-^2 \\ &\leq (\sqrt{n_k}(\overline{X}_{n_k}^* - \tilde{\mu}_{n_k}^0) + c)_-^2 \\ &\rightsquigarrow (\mathbb{Z} + c)_-^2 \text{ as } k \rightarrow \infty \text{ conditional on } \{\tilde{P}_n : n \geq 1\} \\ &\leq (\mathbb{Z})_-^2, \end{aligned} \tag{4.1}$$

meaning that with probability one (with respect to the randomness in  $\{\tilde{P}_n : n \geq 1\}$ ) the EL bootstrap fails to approximate the asymptotic distribution of  $n\widehat{Q}_n(\theta)$  at  $\theta = 0$ .

#### 4.1.2 EL Bootstrap Validity: Modification for the Canonical Example

Next I show that a simple modification to the EL bootstrap yields first order validity. Here I motivate my modification using the previous canonical example where the standard EL bootstrap fails and in the next subsection I discuss the general case. The modification of the EL bootstrap I propose affects both the way the EL probabilities are calculated and the computation of the bootstrap criterion function. Intuitively, instead of imposing  $\mathbb{E}_{P_0}(X - \theta) \geq 0$ , I impose  $\mathbb{E}_{P_0}(X - \theta) \geq \varrho_n$  where  $\varrho_n$  is a sequence of positive random (or deterministic) variables satisfying

$$P_0 \left( \lim \varrho_n = 0 \text{ and } \liminf_{n \rightarrow \infty} \varrho_n (n / (2 \log \log n))^{1/2} \geq 1 \right) = 1. \tag{4.2}$$

**Remark 4.1** When the parameter of interest is on the boundary of the parameter space, Andrews (2000, page 403) proposes a parametric bootstrap procedure in which the parameter estimator used to generate the bootstrap shrinks to the boundary when  $\bar{X}_n$  is below  $\varrho_n$ . The EL bootstrap I propose here does basically the opposite.<sup>19</sup> This is, when  $\theta = 0$  the EL mean is given by  $\bar{\mu}_n^0 = \max\{\bar{X}_n, \varrho_n\}$  so that it keeps the mean far from zero when  $\bar{X}_n$  is too small.

The modified EL bootstrap for this example requires two steps. The first steps involves computing the modified EL probabilities  $\bar{P}_n = (\bar{p}_1, \dots, \bar{p}_n)$  by solving

$$\tilde{l}_{EL} = \sup_{p_1, \dots, p_n} \left\{ \sum_{i=1}^n \log(p_i) \mid p_i > 0; \sum_{i=1}^n p_i = 1; \sum_{i=1}^n p_i (X_i - \theta) \geq \varrho_n \right\}.$$

Once  $\bar{P}_n$  is known, the new empirical likelihood mean is denoted by  $\bar{\mu}_n^\theta = \sum \bar{p}_i (X_i - \theta) = \max\{\bar{X}_n - \theta, \varrho_n\}$ . The second step generates i.i.d. bootstrap samples  $\{X_i^*\}_{i=1}^n$  according to  $\bar{P}_n = (\bar{p}_1, \dots, \bar{p}_n)$  and computes the modified bootstrap criterion function,

$$n\hat{Q}_n^*(\theta) = (\sqrt{n}(\bar{X}_n^* - \theta - \varrho_n))_-^2.$$

Note that by (4.2) and the law of iterated logarithm,

$$P_0 \left( \limsup_{n \rightarrow \infty} (\bar{X}_n - \theta - \varrho_n) \leq 0 \right) = \begin{cases} 0 & \text{if } \theta < 0 \\ 1 & \text{if } \theta = 0 \end{cases}$$

meaning that  $\bar{\mu}_n^\theta > \varrho_n$  for  $n$  large enough with probability one (w.p.1) when  $\theta < 0$  and  $\bar{\mu}_n^\theta = \varrho_n$  for  $n$  large enough w.p.1 when  $\theta = 0$ . Hence,

$$\begin{aligned} n\hat{Q}_n^*(\theta) &= (\sqrt{n}(\bar{X}_n^* - \theta - \varrho_n))_-^2 \\ &= (\sqrt{n}(\bar{X}_n^* - \theta - \bar{\mu}_n^\theta) + \sqrt{n}(\bar{\mu}_n^\theta - \varrho_n))_-^2 \\ &\rightsquigarrow \begin{cases} 0 & \text{if } \theta < 0 \\ (\mathbb{Z})_-^2 & \text{if } \theta = 0 \end{cases} \text{ conditional on } \{\bar{P}_n : n \geq 1\}, \end{aligned}$$

and the modified EL bootstrap is asymptotically valid to approximate the distribution of  $n\hat{Q}_n(\theta)$  for all  $\theta \leq 0$ .

**Remark 4.2** The intuition behind the modified EL bootstrap is as follows. The law of iterated logarithm says that  $(n/(2 \log \log n))^{1/2} \bar{X}_n$  reaches the interval  $(1 - \epsilon, 1 + \epsilon)$  infinitely often w.p.1. This is why the standard EL bootstrap doesn't work in (4.1). When  $\theta = 0$ , the new bootstrap replaces  $\bar{X}_n$  by  $\varrho_n$  which reaches the interval  $(c - \epsilon, c + \epsilon)$  for  $c \geq 1$  infinitely often w.p.1. However, since  $\varrho_n$  is also part of the modified EL bootstrap criterion function, the effect cancels out and the asymptotic distribution is not affected by the behavior of  $\varrho_n$ .

---

<sup>19</sup>A supplementary appendix illustrates the way the modified EL bootstrap works for the example in Andrews (2000)

### 4.1.3 Modified EL Bootstrap Validity: The General Case

Now consider the general case where the economic model imposes  $\mathbb{E}_{P_0}[m(z, \theta)] \geq 0$ . The modified empirical likelihood function is,

$$\tilde{l}_{EL}^r(\theta, \varrho_n^\theta) \equiv \max_{p_1, \dots, p_n} \left\{ \sum_{i=1}^n \log(p_i) \mid p_i \geq 0; \sum_{i=1}^n p_i = 1; \sum_{i=1}^n p_i m(z_i, \theta) \geq \varrho_n^\theta \right\} \quad (4.3)$$

where  $\varrho_n^\theta \geq 0$  is a  $q \times 1$  vector of random (or deterministic) variables that satisfies

$$P_0 \left( \lim \varrho_{j,n}^\theta = 0 \text{ and } \liminf_{n \rightarrow \infty} \varrho_{j,n}^\theta (n / (2 \log \log n))^{1/2} \geq \sqrt{V_{jj}(\theta)} \right) = 1 \quad (4.4)$$

where  $V_{jj}(\theta) = \text{var}_{P_0}\{m_j(z_i, \theta)\}$  for all  $j \in \{1, \dots, q\}$ . By the law of iterated logarithm,

$$P_0 \left( \limsup_{n \rightarrow \infty} (\bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta) \leq 0 \right) = \begin{cases} 0 & \text{if } \mathbb{E}_{P_0}[m_j(z, \theta)] > 0 \\ 1 & \text{if } \mathbb{E}_{P_0}[m_j(z, \theta)] = 0 \end{cases}. \quad (4.5)$$

Denote the modified EL probabilities that solve (4.3) by  $\bar{P}_n = (\bar{p}_1, \dots, \bar{p}_n)$  and define  $\bar{\mu}_n(\theta) = \sum_{i=1}^n \bar{p}_i m(z_i, \theta)$ . Let  $Z_i^*$  be i.i.d. according to  $\bar{P}_n$  so that  $\mathbb{E}_{\bar{P}_n}[m(z^*, \theta)] = \bar{\mu}_n(\theta)$ . The modified bootstrap empirical likelihood function is,

$$\tilde{l}_{EL}^{r*}(\theta, \varrho_n^\theta) \equiv \max_{p_1, \dots, p_n} \left\{ \sum_{i=1}^n \log(p_i) \mid p_i \geq 0; \sum_{i=1}^n p_i = 1; \sum_{i=1}^n p_i (m(z_i^*, \theta) - \varrho_n^\theta) \geq 0 \right\}$$

and the modified bootstrap ELR statistic is,

$$\mathcal{ELR}_n^*(\theta, \varrho_n^\theta) \equiv \max_{\lambda \leq 0} 2 \sum_{i=1}^n \log[1 + \lambda'(m(z_i^*, \theta) - \varrho_n^\theta)]. \quad (4.6)$$

Letting  $N_n$  be the number of bootstrap replications, the  $(1 - \alpha)$ -quantile of  $\{\mathcal{ELR}_{n,k}^*(\theta, \varrho_n^\theta)\}_{k=1}^{N_n}$  is given by,

$$\bar{c}_{n,1-\alpha}^\theta \equiv \inf \left\{ x : \frac{1}{N_n} \sum_{k=1}^{N_n} 1\{\mathcal{ELR}_{n,k}^*(\theta, \varrho_n^\theta) \leq x\} \geq 1 - \alpha \right\}, \quad (4.7)$$

so that the bootstrap confidence region is just,

$$\mathcal{C}_n \equiv \{\theta \in \Theta : \mathcal{ELR}_n(\theta) \leq \bar{c}_{n,1-\alpha}^\theta\}. \quad (4.8)$$

The next theorem shows that the confidence set  $\mathcal{C}_n$  defined in (4.8) satisfies the required coverage condition.

**Theorem 4.1** *Let  $\varrho_n^\theta \geq 0$  be a  $q \times 1$  vector of random (or deterministic) variables satisfying (4.4) and suppose Assumption 3.1 holds. Then,*

I. For all  $\theta \in \Theta_0(P_0)$  the modified bootstrap ELR statistic defined in (4.6) satisfies

$$\mathcal{ELR}_n^*(\theta, \varrho_n^\theta) \rightsquigarrow \bar{\chi}_{b(\theta)}^2(\Sigma_b(\theta))$$

where the convergence is conditional on  $\{\bar{P}_n(\theta) : n \geq 1\}$  for almost every sample path and  $\bar{\chi}_{b(\theta)}^2(\Sigma_b(\theta))$  denotes the chi-bar-square distribution with  $b(\theta)$  terms and covariance matrix  $\Sigma_b(\theta)$ .

II. The confidence set  $\mathcal{C}_n$  defined in (4.8) satisfies

$$\lim_{n \rightarrow \infty} P_0(\theta \in \mathcal{C}_n) \geq 1 - \alpha \quad (4.9)$$

for all  $\theta \in \Theta_0(P_0)$ .

**Proof.** See Appendix A. ■

**Remark 4.3** The statement in (4.9) can be decomposed into two parts. When  $\theta$  is in the boundary of  $\Theta_0(P_0)$  it is the case that  $\lim_{n \rightarrow \infty} P_0(\theta \in \mathcal{C}_n) = 1 - \alpha$ . However, when  $\theta$  is in the interior of  $\Theta_0(P_0)$ ,  $\mathcal{C}_n$  will include  $\theta$  with probability approaching one.

**Remark 4.4** The coverage result in Theorem 4.1 is pointwise consistent in levels. However, a supplementary appendix available upon request shows that the confidence set  $\mathcal{C}_n$  defined in (4.8) also satisfies the type of uniform coverage defined by Andrews and Soares (2007).

**Remark 4.5** Andrews and Soares (2007) recently proved that, given a test statistic, a test based on a generalized moment selection (GMS) critical value has greater power than a subsampling test. The modified EL bootstrap falls into the class of GMS tests and so inherits this property.

Therefore, the modified EL Bootstrap is asymptotically valid up to first order and the confidence region in (4.8) satisfies the coverage requirement. As is the case for the asymptotic distribution and subsampling, the application of this bootstrap procedure is not free of tuning parameters.<sup>20</sup> Although any sequence  $\varrho_n^\theta$  that satisfies the rate in (4.4) will work asymptotically, the finite sample properties of the confidence region might be sensitive to this choice. One advantage here is that there is a value of  $\varrho_n^\theta$  that can serve as a natural benchmark. This value is given by,

$$\bar{\varrho}_{j,n}^\theta \equiv \sqrt{\frac{2 \log \log n \times V_{jj}(\theta)}{n}}, \quad (4.10)$$

which represents a lower bound for  $\varrho_{j,n}^\theta$ , since  $\varrho_{j,n}^\theta = C \times \bar{\varrho}_{j,n}^\theta$  for  $C < 1$  does not satisfy (4.4). There is no reason to believe that  $\bar{\varrho}_n^\theta$  is optimal in any sense but at least the existence of a reference value might ease the choice of  $\varrho_n^\theta$  in applications. Optimal choice of  $\varrho_n^\theta$  is beyond

---

<sup>20</sup>Similar tuning parameters related to law of iterated logarithm are also present in the procedures proposed by Bugni (2007) and Andrews and Soares (2007).

the scope of this paper.<sup>21</sup> In the examples of Section 5 I find that the benchmark value has strong performance, and that coverage appears reasonably robust to this choice.

Before moving to the simulation results, it is worth mentioning that the results from Romano and Shaikh (2008) and Andrews and Guggenberger (2008) can be applied to show that subsampling is a uniformly valid approach to approximate the distribution of  $\mathcal{ELR}_n(\theta)$  and thus confidence regions for  $\theta_0$  can be based on subsampling critical values. The main practical problem in using subsampling lies in choosing the block size  $a_n = o(n)$ . The Monte Carlo simulations of the next section suggest that finite sample coverage of the confidence regions might be more sensitive to the choice of  $a_n$  than the choice of  $\rho_n^\theta$ .

## 5 Monte Carlo Simulations

### 5.1 Missing Data Example

This section takes the setup from Example 1 on missing data to evaluate the finite sample performance of the tools developed in the previous sections. Each simulation experiment depends on three parameters:  $n$ ,  $\alpha$  and  $p$ , where  $n$  is the sample size,  $\alpha$  is the size of the test and  $p$  is the probability of observing  $x$ . I set the propensity score to  $p = 0.7$  and I consider  $n = \{100, 500, 1000\}$  and  $1 - \alpha = \{0.75, 0.85, 0.95\}$ , meaning that the identified set is  $\Theta_0(P_0) = [0.35, 0.65]$ .<sup>22</sup> I then take independent draws of  $x_i \sim \text{Uniform}(0,1)$  and  $w_i \sim \text{Bernoulli}(p)$  to construct the simulated missing data as  $\{(x_i \cdot w_i, w_i)\}_{i=1}^n$ . The number of Monte Carlo replications is equal to 3000.

I compute three empirical likelihood confidence regions that use different critical values. The first one uses an asymptotic critical value. since there is no  $\theta$  such that both moments inequalities are binding simultaneously, the maximum number of binding constraints is one, i.e.  $b^* = 1$ . I can then use the limit distribution in Theorem 3.1 in a straightforward manner without the need of any approximation,

$$\lim_{n \rightarrow \infty} P_0(\mathcal{ELR}_n(\theta) \geq c) = \frac{1}{2}P(\chi_1^2 \geq c), \quad (5.1)$$

so that a confidence region for  $\theta_0$  of level  $1 - \alpha$  simply uses a critical value that solves  $P\{\chi_1^2 \geq c\} = 2\alpha$ . The second confidence region uses subsampling to compute the critical value. I use four different subsample sizes,  $a_n = \{n/10, n/6, n/5, n/4\}$  and in all cases  $N_n = 200$ . Finally, the last set of confidence regions uses the modified empirical likelihood bootstrap introduced in section 4. In this case I set  $N_n = 200$  and use five different values for the parameter  $\varrho_n^\theta = \{\bar{\varrho}_n^\theta, 1.5\bar{\varrho}_n^\theta, 2\bar{\varrho}_n^\theta, 3\bar{\varrho}_n^\theta, 0\}$  where  $\bar{\varrho}_n^\theta$  is the benchmark value defined in (4.10). Notice that the choice  $\varrho_n^\theta = 0$  represents the standard empirical likelihood bootstrap which is inconsistent.

The critical value for each case is computed as  $c_\alpha^* = \sup_{\theta \in [0.35, 0.65]} c_\alpha^\theta$ . This is infeasible in practice but it makes the comparison easier since for each method there is a single critical

<sup>21</sup>Andrews and Jia (2008) recently proposed a way to choose these type of tuning parameters

<sup>22</sup>Different values of  $p$  yield similar results so they are not reported.

value that does not depend on  $\theta$  so that confidence regions are just confidence intervals. This allows me to compare not only coverage but also the length of each confidence interval.

Coverage		0.75			0.85			0.95		
Sample Size		100	500	1000	100	500	1000	100	500	1000
Asymptotic		0.7470	0.7350	0.7323	0.8283	0.8447	0.8307	0.9433	0.9437	0.9400
subsampling	n/10	0.8410	0.8270	0.8067	0.9137	0.9077	0.8610	0.9837	0.9627	0.9277
subsampling	n/6	0.8493	0.8217	0.7743	0.9263	0.9037	0.8403	0.9847	0.9513	0.9044
subsampling	n/5	0.8537	0.8227	0.7580	0.9297	0.8993	0.8283	0.9857	0.9477	0.8944
subsampling	n/4	0.8617	0.8120	0.7477	0.9380	0.8860	0.8087	0.9893	0.9373	0.8794
Bootstrap	$\bar{\varrho}_n^\theta$	0.7573	0.7443	0.7470	0.8377	0.8507	0.8418	0.9477	0.9443	0.9444
Bootstrap	$1.5\bar{\varrho}_n^\theta$	0.7497	0.7483	0.7477	0.8350	0.8513	0.8430	0.9473	0.9460	0.9440
Bootstrap	$2\bar{\varrho}_n^\theta$	0.7507	0.7523	0.7467	0.8320	0.8510	0.8420	0.9473	0.9487	0.9447
Bootstrap	$3\bar{\varrho}_n^\theta$	0.7490	0.7410	0.7470	0.8320	0.8500	0.8420	0.9483	0.9467	0.9453
Bootstrap	0	0.7150	0.6977	0.6963	0.8233	0.8323	0.8243	0.9450	0.9357	0.9420
Length:		0.3			0.3			0.3		
Sample Size		100	500	1000	100	500	1000	100	500	1000
Asymptotic		0.3440	0.3197	0.3136	0.3655	0.3301	0.3209	0.4051	0.3477	0.3338
subsampling	n/10	0.3553	0.3252	0.3140	0.3778	0.3368	0.3227	0.4170	0.3497	0.3315
subsampling	n/6	0.3554	0.3227	0.3028	0.3782	0.3351	0.3165	0.4129	0.3457	0.3273
subsampling	n/5	0.3554	0.3205	0.2993	0.3778	0.3342	0.3123	0.4112	0.3440	0.3249
subsampling	n/4	0.3540	0.3165	0.2951	0.3772	0.3305	0.3076	0.4090	0.3417	0.3206
Bootstrap	$\bar{\varrho}_n^\theta$	0.3458	0.3206	0.3145	0.3678	0.3311	0.3219	0.4076	0.3489	0.3349
Bootstrap	$1.5\bar{\varrho}_n^\theta$	0.3457	0.3208	0.3145	0.3675	0.3312	0.3219	0.4074	0.3491	0.3349
Bootstrap	$2\bar{\varrho}_n^\theta$	0.3452	0.3208	0.3144	0.3668	0.3312	0.3219	0.4068	0.3491	0.3348
Bootstrap	$3\bar{\varrho}_n^\theta$	0.3449	0.3204	0.3145	0.3668	0.3309	0.3218	0.4066	0.3486	0.3348
Bootstrap	0	0.3256	0.3022	0.2978	0.3542	0.3239	0.3147	0.3973	0.3437	0.3316

**Table 1:** Confidence intervals for the identifiable parameter. Missing data with unknown propensity score  $p=0.7$ . 3000 MC replications. 200 bootstrap/subsampling replications.

Table 1 shows the empirical coverage and the length of the confidence intervals for the identifiable parameter. The results show that the modified empirical likelihood bootstrap performs well. A comparison between the bootstrap and subsampling shows that for small sample sizes the bootstrap represents a better approximation. This is even more clear from Table 2 where 9 different values for  $a_n$  were used. For large sample sizes, subsampling does really well in some cases (e.g.,  $n = 1000$  and  $\alpha = 0.25$  with  $a_n = n/4$ ) although its performance seems to be more sensitive to the choice of the block size  $a_n$  than the performance of the bootstrap to the choice of  $\varrho_n^\theta$ .<sup>23</sup> In fact, if we just focus on the benchmark case where  $\varrho_n^\theta = \bar{\varrho}_n^\theta$ , only when  $n = 500$  and  $\alpha = 0.05$  subsampling performs a bit better. Note also that in most cases the modified bootstrap approach performs equally well or even better than the asymptotic approximation. This is worth noticing since the bootstrap requires no knowledge

<sup>23</sup>In a recent finding Bugni (2007) shows that while the rate of approximation of subsampling depends on  $a_n$ , the rate of approximation of his bootstrap approach does not depend on  $\tau_n$ , a sequence which has a similar role than  $\rho_n^\theta$ .



on the number of binding constraints.

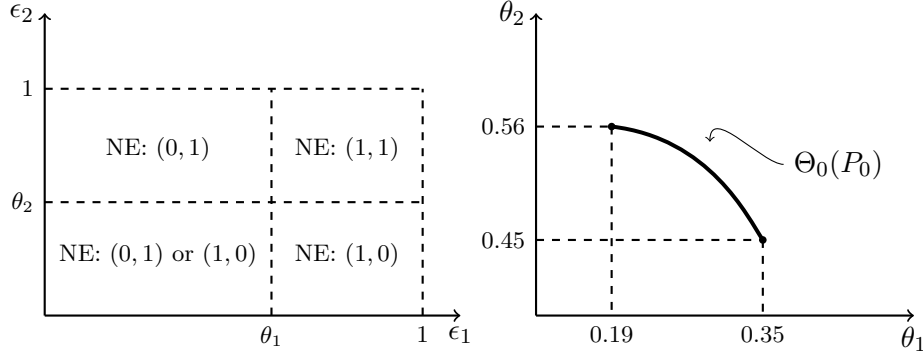
$n = 100$	Asymp.	Bootstrap	Subsampling								
Coverage		$\bar{Q}_n^\theta$	$a_n=10$	$a_n=15$	$a_n=20$	$a_n=25$	$a_n=30$	$a_n=35$	$a_n=40$	$a_n=45$	$a_n=50$
0.75	0.7290	0.7377	0.8310	0.8367	0.8397	0.8503	0.8573	0.8660	0.8727	0.8780	0.8857
0.85	0.8503	0.8563	0.9283	0.9317	0.9410	0.9483	0.9557	0.9633	0.9693	0.9763	0.9793
0.95	0.9513	0.9567	0.9857	0.9857	0.9900	0.9920	0.9953	0.9983	0.9987	0.9990	1.000

**Table 2:** Confidence intervals for the identifiable parameter. Missing data with unknown propensity score  $p=0.7$ . 3000 MC replications. 200 bootstrap/subsampling replications.

Theta	Asymptotic		Subsampling		Asymptotic		Subsampling	
	$\alpha = 0.25$		$\alpha = 0.25$		$\alpha = 0.05$		$\alpha = 0.05$	
	ELR	QLR	ELR	$Q_1$	ELR	QLR	ELR	$Q_1$
0.65	0.2453	0.2428	0.1457	0.1450	0.0532	0.0503	0.0102	0.0098
0.66	0.3520	0.3482	0.2133	0.2107	0.0907	0.0867	0.0192	0.0185
0.67	0.4703	0.4670	0.3045	0.3003	0.1475	0.1400	0.0353	0.0332
0.68	0.5918	0.5892	0.4065	0.4003	0.2250	0.2143	0.0597	0.0567
0.69	0.7082	0.7040	0.5120	0.5090	0.3320	0.3135	0.0980	0.0925
0.70	0.8033	0.8000	0.6305	0.6203	0.4580	0.4362	0.1522	0.1415
0.71	0.8790	0.8760	0.7300	0.7230	0.5847	0.5615	0.2290	0.2068
0.72	0.9348	0.9328	0.8138	0.8043	0.7032	0.6807	0.3197	0.2933
0.73	0.9648	0.9638	0.8807	0.8740	0.8030	0.7838	0.4240	0.3923
0.74	0.9828	0.9822	0.9310	0.9258	0.8837	0.8693	0.5380	0.4967
0.75	0.9932	0.9928	0.9620	0.9587	0.9383	0.9267	0.6457	0.6022

**Table 3:** Finite sample power.

Another interesting property to explore is whether the higher power of the ELR statistic is present in finite sample. In order to analyze this I compute the empirical power (not adjusted for size) at different values of  $\theta$  for three different test statistics including the ELR statistic (see Example 3 for a list). In all cases,  $n = 100$ ,  $\alpha = \{0.25, 0.05\}$ , and  $p = 0.7$  with 6000 replications. The columns labeled Asymptotic compare the performance of the ELR statistic versus the Gaussian quasi-likelihood ratio (QLR) statistic associated with  $Q_2(\theta, P_0)$  in Example 3. Since both of these statistics have the same asymptotic distribution, I use a common critical value from  $P\{\chi_1^2 \geq c\} = 2\alpha$ . The columns under the label Subsampling compare the ELR statistic and the sample analog of  $Q_1(\theta, P_0) \equiv (\mathbb{E}_{P_0}[m(z, \theta)])'_-(\mathbb{E}_{P_0}[m(z, \theta)])_-$  from Example 3. Here, the critical value is computed using subsampling with  $a_n = 25$  and  $N_n = 200$ . The improvements are marginal given the 6000 Monte Carlo replications and so there is no strong evidence that the power of the ELR statistic is above the power of the rest of the alternatives for samples as small as  $n = 100$ .



**Figure 1:** Left: four possible outcomes of the entry game. Right: The identified set when  $\theta_0 = (0.5, 0.3)$ .

## 5.2 Entry Game Example

This section takes the simultaneous entry game from Example 2 where firm  $j \in \{1, 2\}$  has profits  $\pi_{j,m} = (\varepsilon_{j,m} - \theta_j z_{-j,m}) 1\{z_{j,m} = 1\}$  and decides whether to enter ( $z_{j,m} = 1$ ) into a given a market  $m$  or not ( $z_{j,m} = 0$ ). Each simulation experiment depends on three parameters:  $n$ ,  $\alpha$  and  $\theta = (\theta_1, \theta_2)$ . I set  $\theta_0 = (0.5, 0.3)$  and consider  $n = \{500, 1000\}$  and  $\alpha = \{0.85, 0.95\}$ . The four possible Nash Equilibria (NE) situations are presented in Figure 1 (left panel). This model imposes one moment equality<sup>24</sup> and two moment inequalities,

$$\begin{aligned} \mathbb{E}[m_1(z_m, \theta_0)] &= \mathbb{E}[z_{1,m} z_{2,m} - (1 - \theta_1)(1 - \theta_2)] = 0 \\ \mathbb{E}[m_2(z_m, \theta_0)] &= \mathbb{E}[z_{1,m}(1 - z_{2,m}) - \theta_2(1 - \theta_1)] \geq 0 \\ \mathbb{E}[m_3(z_m, \theta_0)] &= \mathbb{E}[\theta_2 - z_{1,m}(1 - z_{2,m})] \geq 0, \end{aligned}$$

where the identified set  $\Theta_0(P_0)$  is given by,

$$\Theta_0(P_0) = \left\{ \theta \in \Theta : \begin{array}{l} \mathbb{E}[z_{1,m}(1 - z_{2,m})] \leq \theta_2 \leq \frac{\mathbb{E}[z_{1,m}(1 - z_{2,m})]}{\mathbb{E}[z_{1,m}(1 - z_{2,m})] + \mathbb{E}[z_{1,m} z_{2,m}]} \\ \theta_1 = 1 - \mathbb{E}[z_{1,m} z_{2,m}] / (1 - \theta_2) \end{array} \right\}.$$

The set  $\Theta_0(P_0)$  has no interior as Figure 1 (right panel) shows for the parameter values used in the simulation.

I simulate the data as follows. I take independent draws of  $\varepsilon_{j,m} \sim U(0, 1)$  for  $j = \{1, 2\}$  and compute the equilibrium according to the region in which  $\varepsilon_m = (\varepsilon_{1,m}, \varepsilon_{2,m})$  falls. If it falls into the region with multiple equilibria, I select  $(1, 0)$  with probability 0.7. Note that in order to simulate the data I need an equilibrium selection mechanism but this information is not used in the inference methods. This is precisely why the model is not point identified.

<sup>24</sup>Introducing moment equalities to the approach proposed in this paper is straightforward: lagrange multipliers associated with moment equalities are unrestricted and those moments have  $\rho_n^\theta = 0$  for the bootstrap implementation.

		n=1000 - coverage=0.85			n=1000 - coverage=0.95		
$\theta_1$	$\theta_2$	Asymp	Boot	SubSam	Asymp	Boot	SubSam
0.3578	0.4550	0.8530	0.8497	0.8027	0.9460	0.9377	0.8793
0.3519	0.4600	0.8893	0.8810	0.8270	0.9653	0.9587	0.8973
0.3458	0.4650	0.9090	0.9037	0.8330	0.9703	0.9677	0.8977
0.3396	0.4700	0.9180	0.9093	0.8317	0.9720	0.9717	0.8983
0.3333	0.4750	0.9200	0.9107	0.8273	0.9727	0.9673	0.8933
0.3269	0.4800	0.9203	0.9083	0.8190	0.9727	0.9653	0.8887
0.3204	0.4850	0.9203	0.8950	0.8140	0.9727	0.9650	0.8830
0.3137	0.4900	0.9203	0.8887	0.8080	0.9727	0.9597	0.8793
0.3069	0.4950	0.9203	0.8770	0.8020	0.9727	0.9540	0.8753
0.3000	0.5000	0.9203	0.8720	0.7997	0.9727	0.9510	0.8713
0.2929	0.5050	0.9203	0.8620	0.7960	0.9727	0.9497	0.8687
0.2857	0.5100	0.9203	0.8560	0.7943	0.9727	0.9437	0.8660
0.2784	0.5150	0.9203	0.8543	0.7907	0.9727	0.9443	0.8660
0.2708	0.5200	0.9203	0.8507	0.7900	0.9727	0.9453	0.8653
0.2632	0.5250	0.9203	0.8557	0.7897	0.9727	0.9433	0.8640
0.2553	0.5300	0.9203	0.8560	0.7897	0.9727	0.9450	0.8657
0.2473	0.5350	0.9203	0.8593	0.7940	0.9727	0.9457	0.8663
0.2391	0.5400	0.9203	0.8703	0.7977	0.9727	0.9517	0.8690
0.2308	0.5450	0.9203	0.8863	0.8020	0.9727	0.9573	0.8757
0.2222	0.5500	0.9203	0.9010	0.8110	0.9727	0.9643	0.8837
0.2135	0.5550	0.9183	0.9093	0.8247	0.9723	0.9697	0.8903
0.2045	0.5600	0.9103	0.9040	0.8243	0.9693	0.9640	0.8933
0.1954	0.5650	0.8570	0.8530	0.7910	0.9520	0.9467	0.8757
Min. Coverage		0.8530	0.8497	0.7897	0.9460	0.9377	0.8640
Ave. Coverage		0.9122	0.8788	0.8070	0.9700	0.9552	0.8788

**Table 4:** Confidence intervals for the entry game. 3000 MC replications. 200 bootstrap/subsampling replications.

Intuitively, each  $\theta$  in the identified set is consistent with some selection mechanism in the region of uncertainty.

As before, I compute three empirical likelihood confidence regions that differ in the way the critical values are computed. The first one uses an asymptotic critical value. In order to deal with the non-pivotalness of the asymptotic distribution, I use the following approximation as in Rosen (2008),

$$\sup_{\theta \in \Theta_0(P_0)} \lim_{n \rightarrow \infty} P_0(\mathcal{ELR}_n(\theta) \geq c) \leq \frac{1}{2} \Pr(\chi_{b^*}^2 \geq c) + \frac{1}{2} \Pr(\chi_{b^*-1}^2 \geq c), \quad (5.2)$$

where  $b^*$  is such that  $\sup_{\theta \in \Theta_0(P_0)} b(\theta) \leq b^*$ . Since there is no  $\theta$  such that all moment inequalities are binding simultaneously,  $b^* = 2$  and I can use (5.2) to compute  $c$  as follows:

$$\frac{1}{2} \Pr(\chi_2^2 \geq c) + \frac{1}{2} \Pr(\chi_1^2 \geq c) = \alpha. \quad (5.3)$$

The second confidence region uses the modified empirical likelihood bootstrap with  $\varrho_n^\theta = \bar{\varrho}_n^\theta$  to compute the critical values. Finally, the last confidence region uses subsampling. Due to computational time constraints I use  $a_n = n/10$  and  $N_n = 200$ .<sup>25</sup>

Table 4 shows the coverage over a grid of values of  $\theta \in \Theta_0(P_0)$  equally spaced between (0.19, 0.56) and (0.35, 0.45) - see Figure 1. As in the missing data example, the results show that the ELR statistic and the modified bootstrap procedure perform well. Note also that at true parameter value  $\theta_0 = (0.3, 0.5)$ , the asymptotic approximation that uses  $b^* = 2$  is too conservative (0.9203) since at that point only one moment is binding. The bootstrap captures this difference and so it gives a better approximation (0.872). The results for the case  $n = 500$  are similar and so are not presented to save space.

## 6 Conclusions

This paper presents results about optimal inference in models that involve moment inequalities and are partially identified. The value of these results lies in the fact that for any of these models there are many different valid criterion functions and each of those leads to a different test statistic, different decision rules and different confidence regions. In this framework, the paper introduces empirical likelihood as a new inference tool for partially identified models and shows that inference based on the empirical likelihood ratio (ELR) statistic is optimal in the so-called Generalized Neyman-Pearson sense. This is, the empirical likelihood ratio statistic is at least as powerful as the rest of the test statistics usually employed for inference in moment inequality models.

Because the asymptotic distribution of the ELR statistic is non-pivotal, this paper also addresses implementation of the new optimal procedure. First I show that the empirical likelihood bootstrap introduced by Brown and Newey (2002) for moment *equalities* does not produce a consistent approximation to the limit distribution of the ELR statistic in the moment *inequality* setting. Then I propose a simple modification of the empirical likelihood bootstrap that recovers first order validity. This modification is simple enough to preserve the computational tricks that typically make EL straightforward to use.

A further inspection of the proposed empirical likelihood framework involves a small scale Monte Carlo simulation. The results show that the modified empirical likelihood bootstrap performs well. The coverage of the confidence regions that employ the new bootstrap approach is very good even for small sample sizes. In addition, the modified bootstrap seems to work well for different values of the tuning parameter and, in particular, the performance using the proposed benchmark value is very satisfactory.

---

<sup>25</sup>Note that 10% of the sample size is a common value used in applications for the block size, see for example Ciliberto and Tamer (2006).

# Appendices

Throughout the appendix let  $m(z_i, \theta)' = [m_b(z_i, \theta)' \ m_s(z_i, \theta)']$  where  $m_b(z_i, \theta)$  is a  $b(\theta) \times 1$  vector of moments with zero mean,  $\mathbb{E}_{P_0}[m_b(z_i, \theta)] = 0$ , and  $m_s(z_i, \theta)$  is a  $s(\theta) \times 1$  vector of moments with positive mean,  $\mathbb{E}_{P_0}[m_s(z_i, \theta)] > 0$ . I use  $\mathbb{E} \equiv \mathbb{E}_{P_0}$  in those cases where  $P_0$  is understood.

## Appendix A Proof of Theorems

**Proof of Theorem 3.1.** Partition  $\lambda(\theta)$  into binding and non-binding moments,  $[\lambda_b(\theta), \lambda_s(\theta)]$  and consider a moment  $j$  such that  $\mathbb{E}[m_j(z_i, \theta)] > 0$ . For such moment  $\exists N$  s.t.  $\forall n \geq N$ ,  $\bar{m}_{j,n}(\theta) > 0$  with probability one (w.p.1). This implies that  $\forall n \geq N$ ,  $\hat{\tau}_{j,n} > 0$  w.p.1 since,

$$\hat{\tau}_{j,n} = \frac{1}{n} \sum_{i=1}^n \left( \frac{m_j(z_i, \theta)}{1 + \hat{\lambda}_n(\theta)' m(z_i, \theta)} \right) \geq \frac{\bar{m}_{j,n}(\theta)}{1 + \hat{\lambda}_n(\theta)' \bar{m}_n(\theta)} \quad (\text{A-1})$$

by Jensen's inequality,  $\hat{\lambda}_n(\theta) \leq 0$  and  $\hat{\lambda}_n(\theta)' \bar{m}_n(\theta) \geq 0$ . The latter follows from

$$0 \leq \max_{\lambda \leq 0} 2 \sum_{i=1}^n \log[1 + \lambda(\theta)' m(z_i, \theta)] \leq 2 \log[1 + \hat{\lambda}_n(\theta)' \sum_{i=1}^n m(z_i, \theta)].$$

Given that  $\hat{\tau}_{j,n} \times \hat{\lambda}_{j,n}(\theta) = 0$  for all  $n$ , it follows that  $\hat{\lambda}_{j,n}(\theta) = 0$  for  $n$  large enough with probability one. We can then conclude that  $\hat{\lambda}_{s,n}(\theta) = 0$  for  $n$  large enough w.p.1. On the other hand, Lemma (B.1) shows that  $\|\hat{\lambda}_{b,n}(\theta)\| = O_p(n^{-1/2})$ . Therefore,  $\exists N$  s.t.  $\forall n \geq N$  the following equalities hold with probability one.

$$\begin{aligned} \mathcal{ELR}_n(\theta) &= \max_{\lambda(\theta) \leq 0} 2 \sum_{i=1}^n \log(1 + \lambda(\theta)' m(z_i, \theta)) \\ &= 2 \sum_{i=1}^n \log(1 + \hat{\lambda}_{b,n}(\theta)' m_b(z_i, \theta) + \hat{\lambda}_{s,n}(\theta)' m_s(z_i, \theta)) \\ &= 2 \sum_{i=1}^n \log(1 + \hat{\lambda}_{b,n}(\theta)' m_b(z_i, \theta)) \\ &= \max_{\lambda_b(\theta) \leq 0} 2 \sum_{i=1}^n \log(1 + \lambda_b(\theta)' m_b(z_i, \theta)) \\ &= \min_{\tau_b \geq 0} \max_{\lambda_b(\theta, \tau_b) \in \mathbb{R}^b} R_n(\theta, \lambda_b, \tau_b). \end{aligned}$$

where  $R_n(\theta, \lambda, \tau)$  is defined in (3.8). Now I can make a similar expansion to that in Owen (1990, pages 100-102) for the expression given above. For an arbitrary sequence  $0 \leq \tau_{b,n} = O_p(n^{-1/2})$ , the first order condition for  $\hat{\lambda}_{b,n}(\theta, \tau_{b,n})$  is given by:

$$- \sum_{i=1}^n \frac{m_b(z_i, \theta)}{(1 + \hat{\lambda}_{b,n}(\theta, \tau_{b,n})' m_b(z_i, \theta))} + n\tau_{b,n} = 0. \quad (\text{A-2})$$

Let  $\gamma_{b,i} \equiv \hat{\lambda}_b(\theta, \tau_{b,n})' m_b(z_i, \theta)$  and note that by Lemmas (B.1) and (B.2) we have:

$$\max_{1 \leq i \leq n} |\gamma_{b,i}| = O_p(n^{-1/2}) o_p(n^{1/2}) = o_p(1). \quad (\text{A-3})$$

Expanding (A-2):

$$\begin{aligned}
0 &= -\sum_{i=1}^n m_b(z_i, \theta) \left(1 - \gamma_i + \frac{\gamma_{b,i}^2}{1 + \gamma_{b,i}}\right) + n\tau_{b,n} \\
&= -\bar{m}_{b,n}(\theta) + \Sigma_{b,n}(\theta) \hat{\lambda}_{b,n}(\theta, \tau_{b,n}) - \frac{1}{n} \sum_{i=1}^n m_b(z_i, \theta) \left(\frac{\gamma_{b,i}^2}{1 + \gamma_{b,i}}\right) + \tau_{b,n} \\
&= -\bar{m}_{b,n}(\theta) + \Sigma_{b,n}(\theta) \hat{\lambda}_{b,n}(\theta, \tau_{b,n}) + \tau_{b,n} + r_{1,n}
\end{aligned}$$

where  $\Sigma_{b,n}(\theta) \equiv n^{-1} \sum_{i=1}^n m_b(z_i, \theta) m_b(z_i, \theta)'$  and  $r_{1,n} \equiv -\frac{1}{n} \sum_{i=1}^n m_b(z_i, \theta) \gamma_{b,i}^2 / (1 + \gamma_{b,i})$  is the reminder term. Then,

$$\hat{\lambda}_{b,n}(\theta, \tau_{b,n}) = \Sigma_{b,n}^{-1}(\theta) (\bar{m}_{b,n}(\theta) - \tau_{b,n}) - \Sigma_{b,n}^{-1}(\theta) r_{1,n} \quad (\text{A-4})$$

where

$$\|r_{1,n}\| \leq \frac{1}{n} \sum_{i=1}^n \|m_b(z_i, \theta)\|^3 \left\| \hat{\lambda}_{b,n}(\theta, \tau_{b,n}) \right\|^2 |1 + \gamma_i|^{-1} = O_p(1) O_p(n^{-1}) O_p(1) = O_p(n^{-1}) \quad (\text{A-5})$$

by Assumption 3.1.

Next, use the fact that  $\log(1 + \gamma_{b,i}) = \gamma_{b,i} - \gamma_{b,i}^2/2 + r_{2,i}$  where for some finite  $C > 0$

$$P(|r_{2,i}| \leq C |\gamma_{b,i}|^3, 1 \leq i \leq n) \rightarrow 1 \quad (\text{A-6})$$

as  $n \rightarrow \infty$ . Now I can approximate the likelihood ratio and then use (A-4).

$$\begin{aligned}
\mathcal{ELR}_n(\theta) &= \min_{\tau_b \geq 0} 2 \sum_{i=1}^n \log(1 + \hat{\lambda}_{b,n}(\theta, \tau_b)' m_b(z_i, \theta)) - 2 \hat{\lambda}_{b,n}(\theta, \tau_b)' n \tau_b \\
&= \min_{\tau_b \geq 0} \left\{ 2 \hat{\lambda}_{b,n}(\theta, \tau_b)' \sum_{i=1}^n m_b(z_i, \theta) - \hat{\lambda}_{b,n}(\theta, \tau_b)' \sum_{i=1}^n m_b(z_i, \theta) m_b(z_i, \theta)' \hat{\lambda}_{b,n}(\theta, \tau_b) \right. \\
&\quad \left. - 2 \hat{\lambda}_{b,n}(\theta, \tau_b)' n \tau_b + 2 \sum_{i=1}^n r_{2,i} \right\} \\
&= \min_{\tau_b \geq 0} \left\{ n \hat{\lambda}_{b,n}(\theta, \tau_b)' (\bar{m}_{b,n}(\theta) - \tau_b) + n \hat{\lambda}_{b,n}(\theta, \tau_b)' r_{1,n} + 2 \sum_{i=1}^n r_{2,i} \right\} \\
&= \min_{\tau_b \geq 0} \left\{ n (\bar{m}_{b,n}(\theta) - \tau_b)' \Sigma_{b,n}^{-1}(\theta) (\bar{m}_{b,n}(\theta) - \tau_b) - n r_{1,n}' \Sigma_{b,n}^{-1}(\theta) r_{1,n} + 2 \sum_{i=1}^n r_{2,i} \right\} \\
&= T_n^\theta + O_p(n^{-1/2}) \quad (\text{A-7})
\end{aligned}$$

where

$$T_n^\theta \equiv \min_{\tau_b \geq 0} \left\{ n (\bar{m}_{b,n}(\theta) - \tau_b)' \Sigma_{b,n}^{-1}(\theta) (\bar{m}_{b,n}(\theta) - \tau_b) \right\} \quad (\text{A-8})$$

$$n r_{1,n}' \Sigma_{b,n}^{-1}(\theta) r_{1,n} = n O_p(n^{-1}) O_p(1) O_p(n^{-1}) = O_p(n^{-1}) \quad (\text{A-9})$$

$$\begin{aligned}
\left| \sum_{i=1}^n r_{2,i} \right| &\leq \sum_{i=1}^n |\gamma_{b,i}|^3 = n \left\| \hat{\lambda}_{b,n}(\theta, \tau_b) \right\|^3 \frac{1}{n} \sum_{i=1}^n \|m_b(z_i, \theta)\|^3 \\
&= n O_p(n^{-3/2}) O_p(1) = O_p(n^{-1/2}). \quad (\text{A-10})
\end{aligned}$$

Next, define

$$\varsigma_{b,n}(\theta) \equiv \sqrt{n} \bar{m}_{b,n}(\theta) \rightarrow^p \varsigma_b(\theta) \sim N(0, \Sigma_b(\theta)) \quad (\text{A-11})$$

so that for all  $\theta \in \Theta_0(P_0)$

$$\begin{aligned} T_n^\theta &= \min_{\tau_b \geq 0} (\varsigma_{b,n}(\theta) - \sqrt{n}\tau_b)' \Sigma_{b,n}^{-1}(\theta) (\varsigma_{b,n}(\theta) - \sqrt{n}\tau_b) \\ &\rightsquigarrow \min_{u_b \geq 0} (\varsigma_b(\theta) - u_b)' \Sigma_b^{-1}(\theta) (\varsigma_b(\theta) - u_b) \end{aligned}$$

where  $u_b = \sqrt{n}\tau_b$  and  $\Sigma_{b,n}(\theta) \rightarrow^p \Sigma_b(\theta)$  by Assumption 3.1.

The statistic  $T^\theta = \min_{u_b \geq 0} (\varsigma_b(\theta) - u_b)' \Sigma_b^{-1}(\theta) (\varsigma_b(\theta) - u_b)$  measures the distance of the normal random variable  $\varsigma_b(\theta) \sim N(0, \Sigma_b(\theta))$  from the nonnegative orthant and by Wolak (1991)

$$P_0 \{T^\theta > c\} = \sum_{j=0}^{b(\theta)} \varpi(b(\theta), b(\theta) - j, \Sigma_b(\theta)) \Pr\{\chi_j^2 \geq c\}.$$

■

**Proof of Theorem 3.2.** The first step in the proof is to show that for each  $\theta \in \Theta$ ,  $\inf_{P \in \mathcal{P}_0(\theta)} I(Q||P)$  is continuous in  $Q$ . For this task I use the following representation. Let  $v_\theta \in \mathcal{V} \subset \mathbb{R}^q$  be such that  $v_\theta = \mathbb{E}_{P_0}[m(z_i, \theta)]$ . Assumptions 3.2 and 3.3 ensure that  $\mathcal{V}$  is compact and then,

$$H_0^\theta : v_\theta \geq 0 \tag{A-12}$$

is equivalent to  $H_0^\theta : \mathbb{E}_{P_0}[m(z_i, \theta)] \geq 0$  and  $\mathcal{P}_0(\theta)$  can be represented as  $\cup_{v \geq 0} \mathcal{P}(\theta, v)$  where

$$\mathcal{P}(\theta, v) \equiv \{P \in \mathcal{M} : \mathbb{E}_P[m(z_i, \theta)] = v\}. \tag{A-13}$$

To simplify notation I omit the dependence on  $\theta$  and note that

$$\inf_{P \in \mathcal{P}_0} I(Q||P) = \inf_{v \geq 0} \inf_{P \in \mathcal{P}(v)} I(Q||P).$$

Let  $q(z) \equiv \frac{dP}{dQ}(z)$  and write  $\inf_{P \in \mathcal{P}(v)} I(Q||P)$  as,

$$\begin{aligned} &\inf_{q(z) \in \{dP/dQ : P \in \mathcal{P}(v)\}} - \int_{\mathcal{Z}} \log q(z) dQ \\ \text{s.t. } \int_{\mathcal{Z}} dP &= \int_{\mathcal{Z}} q(z) dQ = 1, \int_{\mathcal{Z}} (m(z) - v) dP = \int_{\mathcal{Z}} (m(z) - v) q(z) dQ = 0. \end{aligned}$$

By the duality of partially-finite programming (see Borwein and Lewis (1993, Theorem 3.4)), the dual problem is given by

$$\max_{\lambda \in \mathbb{R}^q} \int_{\mathcal{Z}} \log(1 + \lambda'(m(z) - v)) dQ.$$

Due to the concavity of the objective function, the maximizer  $\lambda(Q, v)$  is continuous in  $Q$  by Theorem 10.8 in Rockafeller (1970), which means that

$$\inf_{P \in \mathcal{P}(v)} I(Q||P) = \max_{\lambda \in \mathbb{R}^q} \int_{\mathcal{Z}} \log(1 + \lambda'(m(z) - v)) dQ \tag{A-14}$$

is also continuous in  $Q$ . Compactness of  $\mathcal{Z}$  implies compactness of  $\mathcal{M}(\mathcal{Z})$  so that the function above is uniformly continuous in  $(Q, v) \in \mathcal{M} \times \mathcal{V}$  due to the compactness of  $\mathcal{M} \times \mathcal{V}$ . This implies that the solution  $v(Q)$  to  $\inf_{v \geq 0} \inf_{P \in \mathcal{P}(v)} I(Q||P)$  is also continuous in  $Q$  and the required result follows.

Part (I): The first part of the theorem follows directly from Sanov's Theorem once we realize that continuity of  $\inf_{P \in \mathcal{P}_0(\theta)} I(Q||P)$  implies that the set  $\Lambda_1^\theta = \{Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_0(\theta)} I(Q||P) \geq \eta^\theta\}$  is closed.

Then, Sanov's Theorem implies that for any  $P \in \mathcal{P}_0(\theta)$ ,

$$\begin{aligned} \limsup_{n \rightarrow \infty} n^{-1} \log P^n \{\widehat{P}_n \in \Lambda_1^\theta\} &\leq - \inf_{v \in \Lambda_1^\theta} I(v||P) \\ &\leq -\eta^\theta. \end{aligned}$$

Part (II): Fix  $\theta \in \Theta$ . The second part can be proved following the same arguments used by Kitamura (2001) and Zeitouni and Gutman (1991). The basic idea is to prove that for each  $\theta \in \Theta$  there exists  $n_0 \in \mathbb{N}$  such that

$$\Lambda_0^\theta \subseteq \Omega_{0,n}^\theta \quad (\text{A-15})$$

for all  $n > n_0$ . Suppose it is not so. Then there exists an infinite sequence of measures  $\{\xi_m\}_{m \in \mathbb{N}}$  such that  $\xi_m \in \Lambda_0^\theta$  and  $\xi_m \in \Omega_{1,n_m}^\theta$ . Since the set  $\overline{\Lambda}_0^\theta = \{Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_0(\theta)} I(Q||P) \leq \eta^\theta\}$  is compact in the weak topology (Deuschel and Stroock (1989, Ch.3)), there exists a subsequence  $\{m_k\}_{k \in \mathbb{N}}$  such that  $\xi_{m_k} \rightarrow \xi \in \overline{\Lambda}_0^\theta$ . For such  $\xi$  take the open ball  $B(\xi, \delta/2)$  such that  $B(\xi, \delta/2) \subset (\Omega_{1,n_{m'}}^\theta)^\delta$  for some subsequence  $n_{m'}$ . Moreover, for an  $m^*$  large enough there exists  $\xi_{m^*} \in B(\xi, \delta/2)$  satisfying  $\inf_{P \in \mathcal{P}_0(\theta)} I(\xi_{m^*}||P) < \eta^\theta$ . Therefore,

$$\begin{aligned} \limsup_{n \rightarrow \infty} n^{-1} \log P^n \{\widehat{P}_n \in (\Omega_{1,n}^\theta)^\delta\} &\geq \liminf_{m' \rightarrow \infty} n_{m'}^{-1} \log P^{n_{m'}} \{\widehat{P}_{n_{m'}} \in (\Omega_{1,n_{m'}}^\theta)^\delta\} \\ &\stackrel{(1)}{\geq} \liminf_{m' \rightarrow \infty} n_{m'}^{-1} \log P^{n_{m'}} \{\widehat{P}_{n_{m'}} \in B(\xi, \delta/2)\} \\ &\stackrel{(2)}{\geq} - \inf_{v \in B(\xi, \delta/2)} I(v||P) \\ &\stackrel{(3)}{\geq} -I(\xi_{m^*}||P) \\ &> -\eta^\theta \end{aligned}$$

for all  $P \in \mathcal{P}_0(\theta)$ , where  $\stackrel{(1)}{\geq}$  follows from  $B(\xi, \delta/2) \subset (\Omega_{1,n_{m'}}^\theta)^\delta$ ,  $\stackrel{(2)}{\geq}$  follows from Sanov's theorem and  $\stackrel{(3)}{\geq}$  from  $\xi_{m^*} \in B(\xi, \delta/2)$ . This contradicts

$$\limsup_{n \rightarrow \infty} n^{-1} \log P^n \{\widehat{P}_n \in (\Omega_{1,n}^\theta)^\delta\} \leq -\eta^\theta.$$

Therefore,  $\Lambda_0^\theta \subseteq \Omega_{0,n}^\theta$  for all  $n > n_0$  and this implies,

$$\limsup_{n \rightarrow \infty} n^{-1} \log P_1^n \{\widehat{P}_n \in \Omega_{0,n}^\theta\} \geq \limsup_{n \rightarrow \infty} n^{-1} \log P_1^n \{\widehat{P}_n \in \Lambda_0^\theta\}$$

and

$$\liminf_{n \rightarrow \infty} \frac{P_1^n \{\widehat{P}_n \in \Omega_{0,n}^\theta\}}{P_1^n \{\widehat{P}_n \in \Lambda_0^\theta\}} \geq 1$$

for all  $P_1 \in \mathcal{M}$ . ■

**Proof of Theorem 4.1.** Part (I): Let  $\{\overline{P}_n(\theta) : a.e.\}$  denote “conditional on  $\{\overline{P}_n(\theta) : n \geq 1\}$  for a.e. sample path”. As in the proof of Theorem 3.1 partition  $\lambda(\theta)$  into binding and non-binding moments,  $[\lambda_b(\theta), \lambda_s(\theta)]$  and consider a moment  $j$  such that  $\mathbb{E}[m_j(z_i, \theta)] > 0$ . For such moment,

$$\begin{aligned} \widehat{\tau}_{j,n}^* &= \frac{1}{n} \sum_{i=1}^n \left( \frac{m_j(z_i^*, \theta) - \varrho_{j,n}^\theta}{1 + \widehat{\lambda}_n^*(\theta)'(m(z_i^*, \theta) - \varrho_n^\theta)} \right) \geq \frac{\overline{m}_{j,n}^*(\theta) - \varrho_{j,n}^\theta}{1 + \widehat{\lambda}_n^*(\theta)'(\overline{m}_n^*(\theta) - \varrho_n^\theta)} \\ &= \frac{\overline{m}_{j,n}^*(\theta) - \overline{\mu}_{j,n}(\theta)}{1 + \widehat{\lambda}_n^*(\theta)'(\overline{m}_n^*(\theta) - \varrho_n^\theta)} + \frac{\overline{\mu}_{j,n}(\theta) - \varrho_{j,n}^\theta}{1 + \widehat{\lambda}_n^*(\theta)'(\overline{m}_n^*(\theta) - \varrho_n^\theta)}, \end{aligned} \quad (\text{A-16})$$

where the first term is  $O_p(n^{-1/2})$  conditional on  $\{\overline{P}_n(\theta) : n \geq 1\}$  for almost every sample path and



the second term is positive for  $n$  large enough w.p.1 by Lemma B.4. Thus,  $\widehat{\tau}_{j,n}^* > 0$  for  $n$  sufficiently large  $\{\overline{P}_n(\theta) : a.e.\}$ . Furthermore, since  $\widehat{\tau}_{j,n}^* \times \lambda_{j,n}^* = 0$  for all  $n$ , we can conclude that  $\lambda_{j,n}^* = 0$  for  $n$  sufficiently large  $\{\overline{P}_n(\theta) : a.e.\}$  for all  $j$  such that  $\mathbb{E}[m_j(z_i, \theta)] > 0$ . In addition, by Lemma B.1 we have  $\|\lambda_{b,n}^*\| = O_p(n^{-1/2}) \{\overline{P}_n(\theta) : a.e.\}$ . Therefore, similar arguments as those in the proof of Theorem 3.1 show that we can write  $\mathcal{ELR}_n^*(\theta, \varrho_n^\theta)$  as,

$$\mathcal{ELR}_n^*(\theta, \varrho_n^\theta) \equiv \min_{\tau_b \geq 0} \max_{\lambda_b \in \mathbb{R}^b} 2 \sum_{i=1}^n \log[1 + \lambda_b'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)] - 2n\lambda_b'\tau_b$$

so that the corresponding FOC is,

$$\sum_{i=1}^n \frac{(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)}{(1 + \gamma_{b,i}^*)} - n\tau_b = 0$$

where  $\|\widehat{\lambda}_{b,n}^*\| = O_p(n^{-1/2})$ ,  $\gamma_{b,i}^* \equiv \widehat{\lambda}_{b,n}^{*\prime}(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)$  and  $\max_{1 \leq i \leq n} |\gamma_i^*| = o_p(1)$ . Therefore, using the same expansion used in the proof of theorem 3.1 it is clear that for all  $\theta \in \Theta_0(P_0)$  the statistic  $\mathcal{ELR}_n^*(\theta, \varrho_n^\theta)$  is equivalent to a QLR statistic of the form,

$$\begin{aligned} \mathcal{ELR}_n^*(\theta, \varrho_{b,n}^\theta) &= \min_{\tau_b \geq 0} \left\{ n (\overline{m}_{b,n}^*(\theta) - \varrho_{b,n}^\theta - \tau_b) \widetilde{\Sigma}_{b,n}^{*-1}(\theta) (\overline{m}_{b,n}^*(\theta) - \varrho_{b,n}^\theta - \tau_b) \right\} + o_p(1) \\ &= T_n^{\theta*}(\varrho_{b,n}^\theta) + o_p(1), \end{aligned}$$

where  $\widetilde{\Sigma}_{b,n}^*(\theta) = n^{-1} \sum_{i=1}^n (m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)'$  and  $\overline{m}_{b,n}^*(\theta) = n^{-1} \sum_{i=1}^n m_b(z_i^*, \theta)$ . Now define the random variable  $\varsigma_{b,n}^*(\theta)$  and note that,

$$\varsigma_{b,n}^*(\theta) \equiv \sqrt{n}(\overline{m}_{b,n}^*(\theta) - \overline{\mu}_{b,n}(\theta)) \xrightarrow{p} \varsigma_b(\theta) \sim N(0, \Sigma_b(\theta)) \quad (\text{A-17})$$

where  $\overline{\mu}_{b,n}(\theta) = \sum_{i=1}^n \overline{p}_i m_b(z_i, \theta)$  and the convergence is conditional on  $\{\overline{P}_n(\theta) : n \geq 1\}$  for almost every sample path and for all  $\theta \in \Theta_0(P_0)$ . Then, letting  $u_b = \sqrt{n}(\tau_b - (\overline{\mu}_{b,n}(\theta) - \varrho_{b,n}^\theta))$

$$\begin{aligned} T_n^{\theta*}(\varrho_{b,n}^\theta) &= \min_{\tau_b \geq 0} \left\{ (\varsigma_{b,n}^*(\theta) - \sqrt{n}(\varrho_{b,n}^\theta + \tau_b - \overline{\mu}_{b,n}(\theta))) \widetilde{\Sigma}_{b,n}^{*-1}(\theta) (\varsigma_{b,n}^*(\theta) - \sqrt{n}(\varrho_{b,n}^\theta + \tau_b - \overline{\mu}_{b,n}(\theta))) \right\} \\ &= \min_{u_b \geq -\sqrt{n}(\overline{\mu}_{b,n}(\theta) - \varrho_{b,n}^\theta)} (\varsigma_{b,n}^*(\theta) - u_b) \widetilde{\Sigma}_{b,n}^{*-1}(\theta) (\varsigma_{b,n}^*(\theta) - u_b). \end{aligned}$$

By Lemma B.4, when  $\mathbb{E}_{P_0}[m_j(z, \theta)] = 0$  we have  $\overline{\mu}_{j,n}(\theta) = \varrho_{j,n}^\theta$  for  $n$  large enough w.p.1. Therefore, for the binding moments  $-\sqrt{n}(\overline{\mu}_{b,n}(\theta) - \varrho_{b,n}^\theta) = 0$  with probability 1 for  $n$  sufficiently large, meaning that for  $n$  large enough the following equality holds w.p.1:

$$T_n^{\theta*}(\varrho_{b,n}^\theta) = \min_{u_b \geq 0} (\varsigma_{b,n}^*(\theta) - u_b) \widetilde{\Sigma}_{b,n}^{*-1}(\theta) (\varsigma_{b,n}^*(\theta) - u_b).$$

Since conditional on  $\{\overline{P}_n : n \geq 1\}$ ,  $\widetilde{\Sigma}_{b,n}^*(\theta) \xrightarrow{p} \Sigma_b(\theta)$  for almost every sample path, it follows that  $T_n^{\theta*}(\varrho_{b,n}^\theta) = T^\theta + o_p(1) \{\overline{P}_n(\theta) : a.e.\}$  where

$$T^\theta = \min_{u_b \geq 0} (\varsigma_b(\theta) - u_b)' \Sigma_b(\theta)^{-1} (\varsigma_b(\theta) - u_b). \quad (\text{A-18})$$

This is exactly the same asymptotic distribution of  $\mathcal{ELR}_n(\theta)$ . At this point it is easy to see why the standard EL bootstrap (the one without  $\varrho_n^\theta$ ) does not work. In that case,  $u_b \geq -\sqrt{n}\overline{\mu}_{b,n}(\theta)$  and if we look at a subsequence where  $-\sqrt{n}\overline{\mu}_{b,n}(\theta) \leq -c$  for all  $n$  we have  $u_b \geq -c$  instead of  $u_b \geq 0$ . Such a subsequence exists by the law of iterated logarithm.

Part (II): Since  $T_n^\theta = 0$  wp $\rightarrow$ 1 for any  $\theta$  in the interior of  $\Theta_0(P_0)$ , it follows that for  $\theta \in \text{int}(\Theta_0(P_0))$ ,

$$\liminf_{n \rightarrow \infty} P_0(\mathcal{ELR}_n(\theta) \leq c) = 1$$

for any  $c \geq 0$  and, in particular, for  $\bar{c}_{n,1-\alpha}^\theta$  as defined in (4.7). On the other hand, when  $\theta \in \partial\Theta_0(P_0)$  it follows from the result in Part I that,

$$\liminf_{n \rightarrow \infty} P_0(\mathcal{ELR}_n(\theta) \leq \bar{c}_{n,1-\alpha}^\theta) = 1 - \alpha.$$

■

## Appendix B Auxiliary Lemmas

**Lemma B.1** Consider the set of binding moments,  $\mathbb{E}_{P_0}[m_b(z_i, \theta)] = 0$ , and define  $\hat{\lambda}_{b,n}^*(\theta, \tau_{b,n})$  as,

$$\hat{\lambda}_{b,n}^*(\theta, \tau_{b,n}) \equiv \operatorname{argmax}_{\lambda_b \in \mathbb{R}^b} 2 \sum_{i=1}^n \log[1 + \lambda_b'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)] - 2n\lambda_b'\tau_{b,n} \quad (\text{B-1})$$

where  $z_i^* \sim \bar{P}_n(\theta)$  as defined in (4.3),  $0 \leq \tau_{b,n} = O_p(n^{-1/2})$  and  $\varrho_{b,n}^\theta$  is the sequence in (4.4). Then, under Assumption 3.1,  $\|\hat{\lambda}_{b,n}^*(\theta, \tau_{b,n})\| = O_p(n^{-1/2})$  conditional on  $\{\bar{P}_n : n \geq 1\}$ , for almost every sample path. In addition, letting  $\hat{\lambda}_{b,n}(\theta, \tau_{b,n})$  be defined as in (B-1) but setting  $\varrho_{b,n}^\theta = 0$  and replacing  $z_i^*$  with  $z_i \sim P_0$ , we have  $\|\hat{\lambda}_{b,n}(\theta, \tau_{b,n})\| = O_p(n^{-1/2})$  under the same assumption.

**Proof.** The first order condition for  $\hat{\lambda}_{b,n}^* \equiv \hat{\lambda}_{b,n}^*(\theta, \tau_{b,n})$  is,

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)}{1 + \hat{\lambda}_{b,n}^{*\prime}(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)} - \tau_{b,n} \equiv g(\hat{\lambda}_{b,n}^*).$$

Let  $\hat{\lambda}_{b,n}^* = c_n a_n$ , where  $c_n \geq 0$  and  $\|a_n\| = 1$ . Now,

$$\begin{aligned} 0 &= \|g(c_n a_n)\| \geq |a_n' g(c_n a_n)| \\ &= \frac{1}{n} \left| a_n' \left( \sum_{i=1}^n (m_b(z_i^*, \theta) - \varrho_{b,n}^\theta) - c_n \sum_{i=1}^n \frac{(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta) a_n'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)}{1 + c_n a_n'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)} - \tau_{b,n} \right) \right| \\ &\geq \frac{c_n}{n} \sum_{i=1}^n \frac{a_n'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)' a_n}{1 + c_n a_n'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)} - \frac{1}{n} \left| \sum_{j=1}^b e_j' \sum_{i=1}^n (m_b(z_i^*, \theta) - \varrho_{b,n}^\theta) - \tau_{b,n} \right|, \end{aligned}$$

where  $e_j$  is the unit vector in the  $j$ th coordinate direction. Now let  $A_n = \max_i a_n'(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)$  and note that  $A_n = o(n^{1/2})$  by similar arguments to those in Lemma B.2. Note in addition that  $\tilde{\Sigma}_{b,n}^*(\theta) = n^{-1} \sum_{i=1}^n (m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)(m_b(z_i^*, \theta) - \varrho_{b,n}^\theta)' \xrightarrow{p} \Sigma_b(\theta)$ , conditional on  $\{\bar{P}_n : n \geq 1\}$  for almost every sample path. By Assumption 3.1,  $\sigma_{(1)} > 0$ , where  $\sigma_{(1)}$  is the smallest eigenvalue of  $\Sigma_b(\theta)$  so that  $a_n' \tilde{\Sigma}_{b,n}^*(\theta) a_n \geq \sigma_{(1)} + o_p(1)$  and

$$\begin{aligned} 0 &\geq \frac{c_n a_n' \tilde{\Sigma}_{b,n}^*(\theta) a_n}{1 + c_n A_n} - \frac{1}{n} \left| \sum_{j=1}^b e_j' \sum_{i=1}^n (m_b(z_i^*, \theta) - \varrho_{b,n}^\theta) - \tau_{b,n} \right| \\ &\geq \frac{c_n (\sigma_{(1)} + o_p(1))}{1 + c_n A_n} - \frac{1}{n} \left| \sum_{j=1}^b e_j' \sum_{i=1}^n (m_b(z_i^*, \theta) - \varrho_{b,n}^\theta) - \tau_{b,n} \right| \end{aligned}$$

so that,

$$\begin{aligned} \frac{c_n(\sigma_{(1)} + o_p(1))}{1 + c_n A_n} &\leq \frac{1}{n} \left| \sum_{j=1}^b e'_j \sum_{i=1}^n (m_b(z_i^*, \theta) - \bar{\mu}_{b,n}(\theta)) \right| + \left| \sum_{j=1}^b e'_j (\bar{\mu}_{b,n}(\theta) - \varrho_{b,n}^\theta) \right| + |\tau_{b,n}| \\ &\leq O_p(n^{-1/2}), \end{aligned}$$

conditional on  $\{\bar{P}_n : n \geq 1\}$  for almost every sample path, where  $\bar{\mu}_{b,n}(\theta) = \sum_{i=1}^n \bar{p}_i m_b(z_i, \theta)$ . The last inequality follows because the first term obeys a triangular CLT conditional on the data and the second term satisfies  $\sqrt{n}(\bar{\mu}_{b,n}(\theta) - \varrho_{b,n}^\theta) = 0$  for  $n$  large enough w.p.1 by Lemma B.4. We can then conclude that,

$$c_n = \|\widehat{\lambda}_{b,n}^*(\theta, \tau_{b,n})\| \leq \frac{O_P(n^{-1/2})}{\sigma_{(1)} + o_p(1)}. \quad (\text{B-2})$$

Finally, note that setting  $\varrho_{b,n}^\theta = 0$  and replacing  $z_i^*$  with  $z_i \sim P_0$ , we have  $\|\widehat{\lambda}_{b,n}(\theta, \tau_{b,n})\| = O_p(n^{-1/2})$  following the same steps as above without recentering by  $\bar{\mu}_{b,n}(\theta)$  in the last part. ■

**Lemma B.2** *Let  $z_1, \dots, z_n$  be i.i.d. If  $\sup_{\theta \in \Theta_0(P_0)} \mathbb{E}[\|m(z_i, \theta)\|^\delta] < \infty$  then for all  $\theta \in \Theta_0(P_0)$ ,*

$$P_0 \left\{ \max_{1 \leq i \leq n} \|m(z_i, \theta)\| = o(n^{1/\delta}) \right\} = 1$$

for  $n$  large enough.

**Proof.** This proof follows Owen (1990, lemma 3). Let  $\epsilon > 0$ . Since  $\{z_i\}_{i=1}^n$  is i.i.d. and

$$\sup_{\theta \in \Theta_0(P_0)} \sum_{n=1}^{\infty} P_0 \{ \|m(z_1, \theta)\|^\delta / \epsilon \geq n \} \leq \sup_{\theta \in \Theta_0(P_0)} \mathbb{E}[\|m(z_1, \theta)\|^\delta / \epsilon] < \infty,$$

we have  $\sum_{n=1}^{\infty} P_0\{A_n^\theta\} < \infty$  for all  $\theta \in \Theta_0(P_0)$  where  $A_n^\theta \equiv \{\|m(z_n, \theta)\| \geq \epsilon^{1/\delta} n^{1/\delta}\}$ . By the Borel-Cantelli lemma  $P_0\{A_n^\theta \text{ i.o.}\} = 0$  and this implies  $P_0\{\max_{1 \leq i \leq n} \|m(z_i, \theta)\| \geq \epsilon^{1/\delta} n^{1/\delta} \text{ i.o.}\} = 0$  for all  $\theta \in \Theta_0(P_0)$  so that

$$\limsup_{n \rightarrow \infty} \max_{1 \leq i \leq n} \|m(z_i, \theta)\| / n^{1/\delta} < \epsilon^{1/\delta}$$

with probability 1. Since  $\epsilon$  is arbitrarily small we have that for all  $\theta \in \Theta_0(P_0)$

$$P_0 \left\{ \max_{1 \leq i \leq n} \|m(z_i, \theta)\| = o(n^{1/\delta}) \right\} = 1$$

for  $n$  large enough. ■

**Lemma B.3** *Define the criterion function  $Q_4(\theta, P_0) \equiv \max_{\lambda \leq 0} \mathbb{E}_{P_0}[\log(1 + \lambda' m(z, \theta))]$  and the identified set as,*

$$\Theta_0(P_0) \equiv \{\theta \in \Theta \subseteq \mathbb{R}^k : \mathbb{E}_{P_0}[m(z, \theta)] \geq 0\}.$$

*Then,  $\Theta_0(P_0)$  is equivalent to  $\arg\min_{\theta \in \Theta} Q_4(\theta, P_0)$ .*

**Proof.** First, notice that  $Q_4(\theta, P_0) \geq \mathbb{E}_{P_0}[\log(1)] = 0$ . Fix  $\theta$  and consider  $j \in \{1, \dots, q\}$  such that  $\mathbb{E}_{P_0}[m_j(z, \theta)] > 0$ . Then the multiplier  $\lambda_j$  associated with  $m_j(z, \theta)$  has to be zero. To see this, note that  $\lambda_j < 0$  would induce a contradiction since

$$0 \stackrel{(1)}{=} \mathbb{E}_{P_0} \left[ \frac{m_j(z, \theta)}{1 + \lambda' m(z, \theta)} \right] \stackrel{(2)}{\geq} \frac{\mathbb{E}_{P_0} m_j(z, \theta)}{1 + \lambda' \mathbb{E}_{P_0} m(z, \theta)} > 0$$

where  $\stackrel{(1)}{=}$  follows from the FOC when  $\lambda_j < 0$ ,  $\stackrel{(2)}{\geq}$  follows from Jensen's inequality and the last inequality follows from  $\lambda' \mathbb{E}_{P_0} m(z, \theta) \geq 0$ . Now use the partition  $m(z_i, \theta)' = [m_b(z_i, \theta)' \ m_s(z_i, \theta)']$ . By Jensen's

inequality

$$\begin{aligned}
Q_4(\theta, P_0) &= \max_{\lambda \leq 0} \mathbb{E}_{P_0}[\log(1 + \lambda' m(z, \theta))] \\
&\leq \max_{\lambda \leq 0} \log(1 + \lambda'_s \mathbb{E}_{P_0}[m_s(z, \theta)] + \lambda'_b \mathbb{E}_{P_0}[m_b(z, \theta)]) \\
&= 0,
\end{aligned}$$

since  $\lambda_s = 0$  by the previous argument and  $\mathbb{E}_{P_0}[m_b(z, \theta)] = 0$ . Therefore,  $Q_4(\theta, P_0) = 0$  for all  $\theta \in \Theta_0(P_0)$ . To complete the argument note that  $Q_4(\theta, P_0) > 0$  if  $\theta$  is such that  $\mathbb{E}_{P_0}[m_j(z, \theta)] < 0$  for some  $j \in \{1, \dots, q\}$ . To see this note that  $\mathbb{E}_{P_0}[\log(1 + \lambda' m(z, \theta))]_{\lambda=0} = 0$  and

$$\left. \frac{\partial \mathbb{E}_{P_0}[\log(1 + \lambda' m(z, \theta))]}{\partial \lambda_j} \right|_{\lambda=0} < 0.$$

Let  $\tilde{\lambda}$  be such that  $\tilde{\lambda}_j = -\epsilon$  for some small  $\epsilon > 0$  and  $\tilde{\lambda}_k = 0$  for  $k \neq j$ . By continuity of the objective function in  $\lambda$  we have that  $Q_4(\theta, P_0) \geq \mathbb{E}_{P_0}[\log(1 + \tilde{\lambda}' m(z, \theta))] > \mathbb{E}_{P_0}[\log(1)] = 0$ . ■

**Lemma B.4** Let  $\bar{\mu}_n(\theta) = \sum_{i=1}^n \bar{p}_i m(z_i, \theta)$  denote the Modified EL mean and  $\varrho_n^\theta$  be the sequence in (4.4). Under Assumption 3.1 the following two statements hold.

- (a) If  $\mathbb{E}_{P_0}[m_j(z_i, \theta)] > 0$  then  $\bar{\mu}_{j,n}(\theta) > \varrho_{j,n}^\theta$  for  $n$  large enough with probability one.
- (b) If  $\mathbb{E}_{P_0}[m_j(z_i, \theta)] = 0$  then  $\bar{\mu}_{j,n}(\theta) = \varrho_{j,n}^\theta$  for  $n$  large enough with probability one.

**Proof.** (a) is proved as follows. Recall that  $\bar{\mu}_{j,n}(\theta)$  comes from problem (4.3) and notice that from

$$0 \leq \max_{\lambda \leq 0} 2 \sum_{i=1}^n \log[1 + \lambda'(m(z_i, \theta) - \varrho_n^\theta)] \leq 2 \log[1 + \hat{\lambda}_n' \sum_{i=1}^n (m(z_i, \theta) - \varrho_n^\theta)],$$

it follows that  $\hat{\lambda}_n' \sum_{i=1}^n (m(z_i, \theta) - \varrho_n^\theta) \geq 0$ . Now, from the FOC of  $\lambda_j$ ,

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{m_j(z_i, \theta) - \varrho_{j,n}^\theta}{1 + \hat{\lambda}_n'(m(z_i, \theta) - \varrho_n^\theta)} \right) \stackrel{(1)}{\geq} \frac{\bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta}{1 + \hat{\lambda}_n'(\bar{m}_n(\theta) - \varrho_n^\theta)} \quad (\text{B-3})$$

where  $\stackrel{(1)}{\geq}$  follows from Jensen's inequality. Since  $\hat{\lambda}_n'(\bar{m}_n(\theta) - \varrho_n^\theta) \geq 0$  for all  $n$  and  $\bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta > 0$  for  $n$  large enough w.p.1 by (4.5) we have,

$$P \left( \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left( \frac{m_j(z_i, \theta) - \varrho_{j,n}^\theta}{1 + \hat{\lambda}_n'(m(z_i, \theta) - \varrho_n^\theta)} \right) > 0 \right) = 1$$

meaning that  $\bar{\mu}_{j,n}(\theta) \equiv \frac{1}{n} \sum_{i=1}^n \left( \frac{m_j(z_i, \theta)}{1 + \hat{\lambda}_n'(m(z_i, \theta) - \varrho_n^\theta)} \right) > \varrho_{j,n}^\theta$  w.p.1 for  $n$  sufficiently large.

Now I prove part (b). Set  $\lambda = 0$  and consider the FOC for  $\lambda_j$ ,

$$\left. \frac{1}{n} \sum_{i=1}^n \left( \frac{m_j(z_i, \theta) - \varrho_{j,n}^\theta}{1 + \lambda'(m(z_i, \theta) - \varrho_n^\theta)} \right) \right|_{\lambda=0} = \bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta. \quad (\text{B-4})$$

By (4.5) we know  $\bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta \leq 0$  for  $n$  large w.p.1. If  $\bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta = 0$ , we are done since  $\bar{\mu}_{j,n}(\theta) = \bar{m}_{j,n}(\theta) = \varrho_{j,n}^\theta$  and  $\lambda_j = 0$  is optimal. If  $\bar{m}_{j,n}(\theta) - \varrho_{j,n}^\theta < 0$ , then the optimal value of  $\lambda_j$  has to decrease (so it will be negative) by continuity of the objective function in  $\lambda_j$ . Finally, since the optimal solution has to satisfy  $\lambda_j(\bar{\mu}_{j,n}(\theta) - \varrho_{j,n}^\theta) = 0$  and  $\lambda_j < 0$ , we have  $\bar{\mu}_{j,n}(\theta) = \varrho_{j,n}^\theta$ . ■

## References

- ANDREWS, D. W. K. (2000): “Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space,” *Econometrica*, 68(2), 399–405.
- ANDREWS, D. W. K., S. T. BERRY, AND P. JIA (2004): “Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Locations,” manuscript, Yale University.
- ANDREWS, D. W. K., AND P. GUGGENBERGER (2008): “Validity of Subsampling and ”Plug-in Asymptotic” Inference for Parameters Defined by Moment Inequalities,” *Econometric Theory*, forthcoming.
- ANDREWS, D. W. K., AND P. JIA (2008): “Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure,” manuscript, Yale University.
- ANDREWS, D. W. K., AND G. SOARES (2007): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” manuscript, Yale University.
- BAHADUR, R. (1960): “On the Asymptotic Efficiency of Test and Estimators,” *Sankhya*, 22, 229–252.
- BERESTEANU, A., AND F. MOLINARI (2008): “Asymptotic Properties for a Class of Partially Identified Models,” *Econometrica*, 76(4), 763–814.
- BORWEIN, J. M., AND A. S. LEWIS (1993): “Partially-finite Programing in  $L_1$  and the Existence of Maximum Entropy Estimates,” *SIAM Journal of Optimization*, 3, 248–267.
- BROWN, B. W., AND W. NEWBY (2002): “Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference,” *Journal of Business and Economics Statistics*, 20(4), 507–517.
- BROWN, L. D., G. CASELLA, AND J. T. G. HWANG (1995): “Optimal Confidence Sets, Bioequivalence and the Lamacon of Pascal,” *Journal of the American Statistical Association*, 90(431), 880–889.
- BUGNI, F. (2007): “Bootstrap Inference in Partially Identified Models,” manuscript, Northwestern University.
- CHAMBERLAIN, G. (1987): “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions,” *Journal of Econometrics*, 34(1), 305–334.
- CHERNOFF, H. (1952): “A Measure of Asymptotic Efficiency for Test of a Hypothesis Based on the Sum of Observations,” *The Annals of Mathematical Statistics*, 23, 497–507.
- CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2004): “Inference for Identified Parameter Sets in Econometric Models,” manuscript, M.I.T.
- (2007): “Estimation and Confidence Regions for Parameter Sets in Econometric Models,” *Econometrica*, 75(5), 1243–1284.
- CILIBERTO, F., AND E. TAMER (2006): “Market Structure and Multiple Equilibria in Airline Industry,” manuscript, Northwestern University.
- COHEN, A., AND W. E. STRAWDERMAN (1973): “Admissibility Implications for Different Criteria in Confidence Estimation,” *The Annals of Statistics*, 1(2), 363–366.
- DEMBO, A., AND O. ZEITOUNI (1998): *Large Deviations Techniques and Applications*. Second Edition, Springer: New York.
- DEUSCHEL, J. D., AND D. W. STROOCK (1989): *Large Deviations*. Academic Press.
- DI CICCIO, J. D., P. HALL, AND J. ROMANO (1991): “Empirical Likelihood is Bartlett-Correctable,” *Annals of Statistics*, 19, 1053–1061.
- EVANS, S., B. B. HANSEN, AND P. B. STARK (2005): “Minimax Expected Measure Confidence Sets for Restricted Location Parameters,” *Bernoulli*, 11(4), 571–590.
- GALICHON, A., AND M. HENRY (2006a): “Dilation Bootstrap: A Natural Approach to Inference in Incomplete Models,” manuscript, Harvard University.

- (2006b): “Inference in Incomplete Models,” manuscript, Harvard University.
- GHOSH, J. K. (1961): “On the Relation Among Shortest Confidence Intervals of Different Types,” *Calcutta Statistical Association Bulletin*, 10.
- GOURIEROUX, C., A. HOLLY, AND A. MONFORT (1982): “Likelihood Ratio Test, Wald Test, and Khun-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters,” *Econometrica*, 50(1), 63–80.
- GUGGENBERGER, P., J. HAHN, AND K. KIM (2008): “Specification and Testing Under Moment Inequalities,” *Economics Letters*, 99(2), 375–378.
- HO, K. (2005): “Insurer-Provider Networks in the Medical Care Market,” manuscript, Harvard University.
- HOEFFDING, W. (1965): “Asymptotically Optimal Test for Multinomial Distributions (with Discussion),” *Annals of Mathematical Statistics*, 36, 369–408.
- HOOPER, P. (1982): “Invariant Confidence Sets with Smallest Expected Measure,” *The Annals of Statistics*, 10(4), 1283–1294.
- HOROWITZ, J. L., AND C. F. MANSKI (1998): “Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation using weights and imputations,” *Journal of Econometrics*, 84(1), 37–58.
- (2000): “Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data,” *Journal of the American Statistical Association*, 95(449), 77–84.
- IMBENS, G., AND C. MANSKI (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72, 1845–1857.
- IMBENS, G., R. H. SPADY, AND P. JOHNSON (1998): “Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica*, 66(2), 333–357.
- ISHII, J. (2005): “Interconnection Pricing, Compatibility, and Investment in Network Industries: An Empirical Study of ATM Surcharging in the Retail Banking Industry,” manuscript, Harvard University.
- KITAMURA, Y. (2001): “Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions,” *Econometrica*, 69(6), 1661–1672.
- (2006): “Empirical Likelihood Methods in Econometrics: Theory and Practice,” *Cowles Foundation Discussion Paper*, 1569.
- KITAMURA, Y., AND T. OTSU (2005): “Minimax Estimation and Testing for Moment Condition Models via Large Deviations,” manuscript, Yale University.
- KUDO, A. (1963): “A Multivariate Analog of a One-Sided Test,” *Biometrika*, 59(3), 403–418.
- LEHMANN, E., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. Springer, New York, 3rd edn.
- MADANSKY, A. (1962): “More on Length of Confidence Intervals,” *Journal of the American Statistical Association*, 57(299), 586–589.
- MANSKI, C. (1989): “Anatomy of the Selection Problem,” *Journal of Human Resources*, 24(3), 343–360.
- (2003): *Partial Identification of Probability Distributions*. Springer-Verlag, New York.
- MANSKI, C., AND E. TAMER (2002): “Inference on Regressions with Interval Data on a Regressor or Outcome,” *Econometrica*, 70, 519–546.
- MOON, H. R., AND F. SCHORFHEIDE (2008): “Estimation with Overidentifying Inequality Moment Conditions,” manuscript, University of Pennsylvania.

- NEWHEY, W. K., AND R. J. SMITH (2004): “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72(1), 219–255.
- OWEN, A. (1988): “Empirical Likelihood Ratio Confidence Intervals for a Single Functional,” *Biometrika*, 75, 237–249.
- (1990): “Empirical Likelihood for Confidence Regions,” *Annals of Statistics*, 18(1), 90–120.
- (1991): “Empirical Likelihood for Linear Models,” *Annals of Statistics*, 19(4), 1725–1747.
- OWEN, A. (2001): *Empirical Likelihood*. Chapman and Hall/CRC.
- PAKES, A., J. PORTER, K. HO, AND J. ISHII (2005): “Moment Inequalities and Their Applications,” manuscript, Harvard University.
- PRATT, J. (1961): “Length of Confidence Intervals,” *Journal of the American Statistical Association*, 56(295), 549–567.
- PUHALSKII, A., AND V. SPOKOINY (1998): “On Large-Deviation Efficiency in Statistical Inference,” *Bernoulli*, 4(2), 203–272.
- QIN, B. J., AND J. LAWLESS (1994): “Empirical Likelihood and General Estimating Equations,” *Annals of Statistics*, 22(1), 300–325.
- ROCKAFELLER, R. T. (1970): *Convex Analysis*. Princeton University Press, Princeton.
- ROMANO, J. P., AND A. M. SHAIKH (2006): “Inference for the Identified Set in Partially Identified Econometric Models,” manuscript, Stanford University.
- (2008): “Inference for Identifiable Parameters in Partially Identified Econometric Models,” *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.
- ROSEN, A. (2008): “Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities,” *Journal of Econometrics*, 146(1), 107–117.
- SANTOS, A. (2006): “Inference in Nonparametric Instrumental Variables with Partial Identification,” manuscript, Stanford University.
- SERFLING, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.
- SILVAPULLE, M. J., AND P. K. SEN (2004): *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. Wiley-Interscience, New York.
- SOARES, G. (2006): “Inference for Partially Identified Models with Inequality Moment Constraints,” manuscript, Yale University.
- TAMER, E. (2003): “Incomplete Simultaneous Discrete Response Model with Multiple Equilibria,” *The Review of Economic Studies*, 70, 147–165.
- VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- WOLAK, F. (1987): “An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model,” *Journal of the American Statistical Association*, 82(399), 782–793.
- (1991): “The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models,” *Econometrica*, 59(4), 981–995.
- ZEITOUNI, O., AND M. GUTMAN (1991): “On Universal Hypothesis Testing via Large Deviations,” *IEEE Transactions on Information Theory*, 37(2), 285–290.