EMPIRICAL LIKELIHOOD METHODS IN ECONOMETRICS: THEORY AND PRACTICE

YUICHI KITAMURA*

Abstract. Recent developments in empirical likelihood (EL) are reviewed. First, to put the method in perspective, two interpretations of empirical likelihood are presented, one as a nonparametric maximum likelihood estimation method (NPMLE) and the other as a generalized minimum contrast estimator (GMC). The latter interpretation provides a clear connection between EL, GMM, GEL and other related estimators. Second, EL is shown to have various advantages over other methods. The theory of large deviations demonstrates that EL emerges naturally in achieving asymptotic optimality both for estimation and testing. Interestingly, higher order asymptotic analysis also suggests that EL is generally a preferred method. Third, extensions of EL are discussed in various settings, including estimation of conditional moment restriction models, nonparametric specification testing and time series models. Finally, practical issues in applying EL to real data, such as computational algorithms for EL, are discussed. Numerical examples to illustrate the efficacy of the method are presented.

1. Introduction

Likelihood-based methods are of fundamental importance in econometrics. When the model is correctly specified, the maximum likelihood (ML) procedure automatically yields an estimator that is asymptotically efficient in several senses. For instance, the maximum likelihood estimator (MLE) is a best asymptotically normal (BAN) estimator (see, for example, Chapter 4 of Serfling (1980)) under regularity conditions. It is known that a bias corrected MLE is higher order efficient (Ghosh (1994)). Other concepts of asymptotic efficiency also point to the superiority of MLE. For example, consider the following asymptotic efficiency criterion in terms of "large deviations" (see Chapter 10 of Serfling (1980) for discussions on the large deviation theory). Suppose a random sample $(z_1, ..., z_n)$ is generated

Date: This Version: May 4, 2006.

Keywords: Conditional moment restriction, Empirical distribution, GMM, Large deviation principle, Moment restriction models, Nonparametric specification test, NPMLE, Weak dependence.

JEL Classification Number: C14.

[*] I acknowledge financial support from the National Science Foundation via grant SES-0241770. This paper is prepared for an invited symposium at the 9th World Congress of the Econometric Society in London. I thank Whitney Newey, Richard Smith and participants of the World Congress for valuable comments.

according to a parametric probability measure P_θ. It is known that, in general, the probability of a consistent estimator $\theta_n = \theta_n(z_1, ..., z_n)$ missing its true value θ by a margin exceeding a fixed value c decays exponentially as n goes to infinity. The (negative of the) decay rate

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\theta\{\|\theta_n - \theta\| > c\}, \quad c > 0 \tag{1.1}$$

has been used to measure the efficiency of θ_n. Obviously, an estimator that makes the "rate" (1.1) small is desirable. Kester and Kallenberg (1986) show that MLE achieves the lower bound of the above rate if the parametric model belongs to the convex exponential family. The last requirement is rather restrictive, but it is removable in the sense that MLE is generally optimal if the limit of the rate (1.1) as $c \to 0$ is used as an efficiency criterion; see Bahadur (1960) and Bahadur, Zabell, and Gupta (1980).

Inference methods based on likelihood also possess a number of desirable properties. A leading example is the celebrated Neyman-Pearson Fundamental Lemma. Moreover, the large deviation principle (LDP) uncovers further optimality properties of the likelihood ratio test in broader contexts. Hoeffding (1963) considers a multinomial model and shows that the likelihood ratio test is optimal in terms of large deviation probabilities of type II errors. This optimality of the likelihood ratio test has been extended to more general hypothesis testing problems for parametric distributions (Zeitouni and Gutman (1991)).

As widely recognized, the validity of the likelihood approach generally depends on the assumption on the parametric form for the data distribution, and this fact has spurred the development of nonparametric and semiparametric methods. Perhaps one of the earliest ideas of treating the data distribution nonparametrically in statistical estimation and testing is to use the empirical distribution of the data by comparing it with the (family of) distribution(s) implied by a statistical model. This requires some measure of divergence between distributions. Standard testing methods such as the Kolmogorov-Smirnov test fall into this category, but the estimation theory based on the idea has been developed as well, as exemplified by the classic treatise by Wolfowitz (1957) on the minimum distance estimation. See Manski (1983) as well as Brown and Wegkamp (2002) for further developments of this line of research in econometrics. An estimation procedure that generalizes the minimum distance method by Wolfowitz is studied by Bickel, Klassen, Ritov, and Wellner (1993), who call it the Generalized Minimum Contrast (GMC) method; see Section 3 for more discussion on GMC. The minimum contrast approach yields procedures that are robust against distribution assumptions, though potentially at the cost of efficiency.

It has been recognized that the notion of likelihood can be introduced in the empirical minimum contrast framework just described above. This raises a conjecture: by using likelihood as a measure of distance, it may be possible to develop a method that is robust against distributional assumptions yet possesses good properties analogous to that of a parametric likelihood procedure. Remarkably, recent research shows that this conjecture holds, at least for certain classes of models that are important in econometrics. In particular, this idea yields a powerful and elegant procedure when applied to moment condition models. In his important paper, Owen (1988) has coined term "empirical likelihood" for this procedure. Its literature has been growing rapidly since then, as documented in Owen (2001). The current paper illustrates the method by connecting it with two important existing statistical frameworks, one being nonparametric MLE (NPMLE) and the other GMC. It also gives an updated review of the literature and provides some practical guidance for applied econometricians.

2. EL as NPMLE

This section treats empirical likelihood as a nonparametric maximum likelihood estimation procedure (NPMLE). The basic idea of NPMLE is simple. Suppose the econometrician observe IID data $\{z_i\}_{i=1}^n$, where each z_i is distributed according to an unknown probability measure μ. The fundamental concept is the nonparametric (or empirical) log likelihood function:

$$\ell_{\mathrm{NP}}(p_1, ..., p_n) = \sum_{i=1}^n \log p_i, \text{ where } \sum_{i=1}^n p_i = 1. \tag{2.1}$$

This can be interpreted as the log likelihood for a multinomial model, where the support of the multinomial distribution is given by the empirical observations $\{z_i\}_{i=1}^n$, even though the distribution μ of z_i is not assumed to be multinomial. Rather, μ is left unspecified and therefore it is treated nonparametrically. It is obvious that the maximum of the above log-likelihood function is attained at $p_i = \frac{1}{n}$, therefore the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ (δ_z denotes a unit mass at z) can be regarded as the nonparametric maximum likelihood estimator (NPMLE) for the unknown probability measure μ. The maximum value of ℓ_{NP} is $-n \log n$. See Bickel, Klassen, Ritov, and Wellner (1993) (Section 7.5 in particular) for a more rigorous derivation of the empirical measure μ_n as a NPMLE for μ.

The above example involves no model, but NPMLE works for well-specified econometric models as well. Owen (1990) made a crucial observation that the nonparametric maximum likelihood method shares many properties with conventional parametric likelihood when applied to moment condition models. Consider the model

$$E[g(z_i,\theta)] = \int g(z,\theta)d\mu = 0, \theta \in \Theta \subset \mathbb{R}^k. \tag{2.2}$$

where g is a known $\mathbb{R}^q$-valued function. The unknowns in the above model are θ and μ. The symbols θ_0 and μ_0 are used to denote the true values of θ and μ.

To apply NPMLE to (2.2), "parameterize" the model by $(\theta, p_1, ..., p_n)$ that resides in $\Theta \times \Delta$, where Δ denotes the simplex $\{(p_1, ..., p_n) : \sum_{i=1}^n p_i = 1, 0 \leq p_i, i = 1, ..., n\}$. The nonparametric log-likelihood function to be maximized is

$$\ell_{\mathrm{NP}} = \sum_{i=1}^n \log p_i, \quad \sum_{i=1}^n g(z_i,\theta)p_i = 0.$$

The value of $(\theta, p_1, ..., p_n) \in \Theta \times \Delta$ that maximizes ℓ_{NP} is called the (maximum) empirical likelihood estimator and denoted by $(\hat{\theta}_{\mathrm{EL}}, \hat{p}_{\mathrm{EL}1}, ..., \hat{p}_{\mathrm{EL}n})$. The NPMLE for θ_0 and μ_0 are $\hat{\theta}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{EL}} = \sum_{i=1}^n \hat{p}_{\mathrm{EL}i}\delta_{z_i}$. One might expect that the high dimensionality of the parameter space $\Theta \times \Delta$ makes the above maximization problem difficult to solve for any practical application. Fortunately, that is not the case. Instead of maximizing ℓ_{NP} with respect to the parameters $(\theta, p_1, ..., p_n)$ jointly, first fix θ at a given value of θ and consider the log-likelihood with the parameters $(p_1, ..., p_n)$ "profiled out":

$$\ell(\theta) = \max \ell_{\mathrm{NP}}(p_1, ..., p_n) \text{ subject to } \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g(z_i,\theta) = 0. \tag{2.3}$$

Once this is done, maximize the profile likelihood $\ell(\theta)$ to obtain the empirical likelihood estimator. It turns out that (2.3) is easy to solve numerically, as illustrated below.

The Lagrangian associated with the constrained optimization problem (2.3) is

$$\mathcal{L} = \sum_{i=1}^n \log p_i + \lambda(1 - \sum_{i=1}^n p_i) - n\gamma' \sum_{i=1}^n p_i g(z_i,\theta),$$

where $\lambda \in \mathbb{R}$ and $\gamma \in \mathbb{R}^q$ are Lagrange multipliers. It is a straightforward exercise to show that the first order conditions for $\mathcal{L}$ are solved by:

$$\hat{\lambda} = n, \quad \hat{\gamma}(\theta) = \operatorname*{argmin}_{\gamma \in \mathbb{R}^q} - \sum_{i=1}^n \log(1 + \gamma' g(z_i,\theta))), \qquad \text{and}$$

$$\hat{p}_i(\theta) = \frac{1}{n(1 + \hat{\gamma}(\theta)' g(z_i,\theta))}, \tag{2.4}$$

yielding

$$\ell(\theta) = \min_{\gamma \in \mathbb{R}^q} - \sum_{i=1}^{n} \log(1 + \gamma' g(z_i, \theta))) - n \log n. \tag{2.5}$$

The empirical likelihood estimator for θ_0 is therefore

$$\hat{\theta}_{\mathrm{EL}} = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta) = \operatorname*{argmax}_{\theta \in \Theta} \min_{\gamma \in \mathbb{R}^q} - \sum_{i=1}^{n} \log(1 + \gamma' g(z_i, \theta)).$$

The numerical evaluation of the function $\ell(\cdot)$ is easy, because (2.5) is a low dimensional convex maximization problem, for which a simple Newton algorithm works. The maximization of $\ell(\theta)$ with respect to θ is typically carried our using a nonlinear optimization algorithm. Once $\hat{\theta}_{\mathrm{EL}}$ is calculated, $\hat{p}_{\mathrm{EL}i}, i = 1, ..., n$ are obtained using the formula (2.4):

$$\hat{p}_{\mathrm{EL}i} = \frac{1}{n(1 + \hat{\gamma}(\hat{\theta}_{\mathrm{EL}})' g(z_i, \hat{\theta}_{\mathrm{EL}}))}. \tag{2.6}$$

More computational issues will be discussed in Section 8.

Qin and Lawless (1994) derived the asymptotic distribution of the empirical likelihood estimator. Let $D = E[\nabla_\theta g(z, \theta_0)]$ and $S = E[g(z, \theta_0) g(z, \theta_0)']$, then

$$\sqrt{n}(\hat{\theta}_{\mathrm{EL}} - \theta_0) \xrightarrow{d} \mathrm{N}(0, (D' S D)^{-1}). \tag{2.7}$$

The asymptotic variance coincides with the semiparametric efficiency bound derived by Chamberlain (1987). (Note that Chamberlain (1987) also uses a sequence of approximating multinomial models in his argument.) It is interesting to observe that maximizing the nonparametric likelihood function ℓ_{NP} for the moment condition model (2.2) automatically achieves efficiency. This is a semiparametric analog of the standard result that maximizing the likelihood function of a parametric model yields an efficient estimator. The estimator $\hat{\mu}_{\mathrm{EL}} = \sum_{i=1}^{n} \hat{p}_{\mathrm{EL}i} \delta_{z_i}$ is also an efficient estimator for μ in the following sense. Suppose one wishes to estimate the expectation of a function $a(z, \theta_0)$ of z, i.e. $E(a(z, \theta_0)) = \int a(z, \theta_0) d\mu$. Using $\hat{\mu}_{\mathrm{EL}}$, let $E(\widehat{a(z, \theta_0)}) = \int a(z, \hat{\theta}_{\mathrm{EL}}) d\hat{\mu}_{\mathrm{EL}} = \sum_{i=1}^{n} \hat{p}_{\mathrm{EL}i} a(z_i, \hat{\theta}_{\mathrm{EL}})$. This estimator is more efficient than a naive sample mean such as $\frac{1}{n} \sum_{i=1}^{n} a(z_i, \hat{\theta}_{\mathrm{EL}})$, and can be shown to be semiparametrically efficient, using a result obtained by Brown and Newey (2002).

Empirical likelihood also applies to testing problems. Let R denote a known $\mathbb{R}^s-$valued function of θ. Suppose the econometrician poses a hypothesis that θ_0 is restricted as $R(\theta_0) = 0$ (and assume that the s restrictions are independent). This can be tested by forming a nonparametric analog of the parametric likelihood ratio statistic

$$r = -2\left(\sup_{\theta:R(\theta)=0} \ell(\theta) - \sup_{\theta\in\Theta} \ell\right) \tag{2.8}$$
$$= -2\left(\sup_{\theta:R(\theta)=0} \ell(\theta) - \ell(\hat{\theta}_{\mathrm{EL}})\right),$$

which obeys the chi-square distribution with s degrees of freedom asymptotically under the null that $R(\theta_0) = 0$. This is called the empirical likelihood ratio (ELR) statistic. Another interesting possibility is to define the empirical log likelihood ratio *function*

$$\mathrm{elr}(\theta) = -2\left[\ell(\theta) - (-n\log n)\right] = \max_{\gamma\in\mathbb{R}^q} 2\sum_{i=1}^n \log(1+\gamma' g(z_i,\theta))). \tag{2.9}$$

The first and the second terms in the square bracket are the maximized values of the log nonparametric likelihood with and without the restriction $\sum_{i=1}^n p_i g(z_i,\theta) = 0$, respectively. It can be shown that its value at θ_0, i.e. $\mathrm{elr}(\theta_0)$, obeys the χ^2_q distribution asymptotically under (2.2) and mild regularity conditions; see Owen (1991) and Section 3.5 of Owen (2001). Note that this procedure tests the overidentifying restrictions (2.2) and the parametric restriction $\theta = \theta_0$ jointly, since the restriction $\sum_{i=1}^n p_i g(z_i,\theta_0) = 0$ imposes the two restrictions simultaneously. Thus it is similar to the Anderson-Rubin test (Anderson and Rubin (1949)) in its scope. Finally, if one wants to test the overidentifying restrictions only, the restricted log likelihood in (2.9) is maximized under the constraint (2.2) but treating θ as a free parameter, therefore the corresponding restricted and the unrestricted empirical log-likelihood are $\ell(\hat{\theta}_{\mathrm{RL}})$ and $-n\log n$, respectively. The empirical likelihood ratio statistic for the overidentification hypothesis (2.2), therefore, is $\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})$. This statistic obeys the chi-square distribution with $q-k$ degrees of freedom asymptotically.

Some may find having various versions of empirical likelihood ratio statistics rather confusing. The following elementary relationships among the statistics might help to clarify this. Suppose one wishes to test a parametric hypothesis of the form $\theta = \theta_0$. Then the appropriate statistic is $r = -2\left(\ell(\theta_0) - \ell(\hat{\theta}_{\mathrm{EL}})\right)$, which is equal to

$$\begin{aligned} r &= -2\left(\ell(\theta_0) - \ell(\hat{\theta}_{\mathrm{EL}})\right) \\ &= \left[-2\left(\ell(\theta_0) + n\log n\right)\right] - \left[-2\left(\ell(\hat{\theta}_{\mathrm{EL}}) + n\log n\right)\right] \\ &= \mathrm{elr}(\theta_0) - \mathrm{elr}(\hat{\theta}_{\mathrm{EL}}), \end{aligned}$$

or

$$\mathrm{elr}(\theta_0) = r + \mathrm{elr}(\hat{\theta}_{\mathrm{EL}}). \tag{2.10}$$

This is similar to the decomposition noted in, for example, Stock and Wright (2000) (page 1066). The last equation shows that the test statistic $\mathrm{elr}(\theta_0)$, which tests the k parametric hypotheses and $q-k$ overidentifying restrictions simultaneously, splits into the empirical likelihood ratio test statistic for $\theta = \theta_0$ and the empirical likelihood-based test of the overidentifying restrictions.

A moment condition model is a prime example for which nonparametric maximum likelihood works very well. Note, however, that NPMLE has been applied to other models. For example, Cosslett (1983) considers a binary choice model

$$y_i = 1\{x_i'\theta + \epsilon_i > 0\}, \theta \in \Theta \subset \mathbb{R}^k$$

where ϵ_i is independent of x_i. Here the unknown parameters are the finite dimensional parameter θ and the distribution of ϵ. To put Cosslett's estimator in our framework, consider a probability measure for ϵ that puts probability mass of p_i on each $\{-x_i'\theta\}, i = 1, ..., n$. Then the empirical log likelihood (or the nonparametric log likelihood) for $(\theta, p_1, ..., p_n)$ is given by

$$\ell_{\mathrm{NP}} = \sum_{i=1}^n \left[y_i \log \left(\sum_{j=1}^n 1\{x_j'\theta \le x_i'\theta\} p_i \right) + (1 - y_i) \log \left(1 - \sum_{j=1}^n 1\{x_j'\theta \le x_i'\theta\} p_i \right) \right].$$

Maximizing this empirical likelihood function for $(\theta, p_1, ..., p_n)$ over $\Theta \times \Delta$ yields Cosslett's estimator. Many other applications of NPMLE have been considered in the econometrics literature, e.g. Heckman and Singer (1984); see also Cosslett (1997).

3. EL as GMC

3.1. **GMC and Duality.**

This section offers an interpretation of empirical likelihood alternative to the one as a nonparametric ML procedure given in the previous section. As noted by Bickel, Klassen, Ritov, and Wellner (1993), it is useful to cast (parametric) MLE as a special case of the generalized minimum contrast (GMC) estimation procedure. This principle can be applied here to construct a family of estimators to which EL belongs as a special case. Consider a contrast function that measures the divergence between two probability measures P and Q:

$$D(P, Q) = \int \phi \left(\frac{dP}{dQ} \right) dQ, \tag{3.1}$$

where ϕ is chosen so that it is convex. If P is not absolutely continuous with respect to Q, define the divergence D to be ∞. $D(\cdot, P)$ is minimized at P.

The econometrician observes IID draws of an $\mathbb{R}^p-$valued random variable z that obeys the probability measure μ, and considers the model of the form (2.2). To interpret EL as a version of GMC, introduce the following notation. Let $\mathbf{M}$ denote the set of all probability measures on $\mathbb{R}^p$ and

$$\mathcal{P}(\theta) = \left\{ P \in \mathbf{M} : \int g(z,\theta) dP = 0 \right\}.$$

Define

$$\mathcal{P} = \cup_{\theta \in \Theta} \mathcal{P}(\theta), \tag{3.2}$$

which is the set of all probability measures that are compatible with the moment restriction (2.2). The set $\mathcal{P}$ is called a statistical model. It is correctly specified if and only if $\mathcal{P}$ includes the true measure μ as its member. At the population level, the GMC optimization problem is:

$$\inf_{P \in \mathcal{P}} D(P,\mu) = \inf_{\theta \in \Theta} \inf_{P \in \mathcal{P}(\theta)} D(P,\mu). \tag{3.3}$$

If the model is correctly specified, the minimum is attained by $P = \mu$ in the first expression and $\theta = \theta_0$ in the second expression. (3.3) is a variational problem as the minimization problem $\inf_{P \in \mathcal{P}(\theta)} D(P,\mu)$ involves optimization over functions. Using a variational problem as a basis of estimation may seem unpractical from a computational point of view. Fortunately, a duality theorem in the convex analysis comes to rescue. For a value θ in Θ, consider the infinite dimensional constrained optimization problem

$$v(\theta) = \inf_P D(P,\mu) = \inf_P \int \phi\left(\frac{dP}{d\mu}\right) d\mu \tag{\mathbf{P}}$$

$$\text{subject to} \int g(z,\theta)dP = 0, \int dP = 1,$$

where $v(\theta)$ is the value function corresponding a particular choice of θ. The primal problem $(\mathbf{P})$ has a dual problem

$$v^\star(\theta) = \max_{\lambda \in \mathbb{R}, \gamma \in \mathbb{R}^q} \left[\lambda - \int \phi^\star(\lambda + \gamma' g(z,\theta)) d\mu \right], \tag{\mathbf{DP}}$$

where $\phi^\star$ is the convex conjugate (or the Legendre transformation) of ϕ;[1] see Borwein and Lewis (1991). Note $(\mathbf{DP})$ is a finite dimensional unconstrained convex maximization problem.

[1] For a convex function $f(x)$, its convex conjugate $f^\star$ is given by

$$f^*(y) = \sup_x \left[xy - f(x) \right].$$

The Fenchel duality theorem (see Borwein and Lewis (1991)) implies that[2]

$$v(\theta) = v^\star(\theta). \tag{3.4}$$

Let $(\phi')^{-1}$ denote the inverse of the derivative of ϕ, then the minimum of ($\mathbf{P}$) is attained by $P = \bar{P}$ such that:

$$d\bar{P}(\theta) = (\phi')^{-1}(\lambda + \gamma' g(z,\theta))d\mu. \tag{3.5}$$

See Borwein and Lewis (1991) for details. Equations (3.3), ($\mathbf{P}$), ($\mathbf{DP}$) and (3.4) show that θ_0 solves the minimization problem

$$\inf_{\theta\in\Theta} v^\star(\theta) = \inf_{\theta\in\Theta} \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q} \left[\lambda - \int \phi^\star(\lambda + \gamma' g(z,\theta))d\mu\right]. \tag{3.6}$$

The preceding discussion focused on the population. Statistical procedures can be obtained by replacing the unknown μ with the empirical measure μ_n. By (3.1) and (3.3), an appropriate sample version of the GMC minimization problem is

$$\inf_{P\in\mathcal{P},P\ll\mu_n} \frac{1}{n}\sum_{i=1}^n \phi(np_i) = \inf_{\theta\in\Theta} \inf_{P\in\mathcal{P}(\theta),P\ll\mu_n} \frac{1}{n}\sum_{i=1}^n \phi(np_i), \tag{3.7}$$

where p_i denotes the probability mass that P puts on each point x_i and $P \ll \mu_n$ means that P is absolutely continuous with respect to μ_n. Equation (3.7) leads to the following definition of the GMC estimator for θ:

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \widehat{v}(\theta), \quad \widehat{v}(\theta) = \inf_{\substack{\sum_{i=1}^n p_i g(z_i,\theta)=0 \\ \sum_{i=1}^n p_i = 1}} \frac{1}{n}\sum_{i=1}^n \phi(np_i). \tag{3.8}$$

The formulation based on the sample version of the GMC problem corresponds to the use of "empirical discrepancy statistics" by Corcoran (1998). See also Kitamura (1996b), where its is noted that the discrepancy measure $D(P,\mu) = \int \phi(\frac{dP}{d\mu})d\mu$ is essentially the $f-$divergence by Csiszàr (1967). Newey and Smith (2004) refer to the sample GMC-based estimator as the minimum distance estimator.

The duality theorem shows that (3.8) is equivalent to a computationally convenient form

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \widehat{v}^\star(\theta), \quad \widehat{v}^\star(\theta) = \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q} \left[\lambda - \frac{1}{n}\sum_{i=1}^n \phi^\star(\lambda + \gamma' g(z_i,\theta))\right], \tag{3.9}$$

obtained by replacing μ with μ_n in (3.6). This also yields a natural estimator for μ. The expression (3.5) with μ replaced by μ_n yields

$$\widehat{\bar{P}(\hat{\theta})}(A) = \int_A (\phi')^{-1}(\lambda + \gamma' g(z,\hat{\theta}))d\mu_n \tag{3.10}$$

[2] The nonnegativity of $\frac{dP}{d\mu}$ is guaranteed if ϕ is modified appropriately as in Borwein and Lewis (1991).

as an estimator for $\mu(A)$ for every Borel set A defined on the sample space of z_i.

Choosing $\phi(x)$ to be $-\log(x)$ corresponds to empirical likelihood, because letting $\phi(x) = -\log(x)$ in (3.8) yields a GMC estimator of the form:

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \inf_{\substack{\sum_{i=1}^n p_i g(z_i,\theta)=0 \\ \sum_{i=1}^n p_i=1}} \frac{1}{n}\sum_{i=1}^n -\log(np_i), \tag{3.11}$$

which is exactly the definition of the empirical likelihood estimator given in Section 2. Note that the convex conjugate of $\phi(x) = -\log(x)$ is $\phi^\star(y) = -1 - \log(-y)$. Using this expression in (3.9) and concentrating λ out, obtain

$$\begin{aligned} \hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmin}} \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q} \left[\lambda + 1 + \frac{1}{n}\sum_{i=1}^n \log(-\lambda - \gamma' g(z_i,\theta))\right] \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} \max_{\gamma\in\mathbb{R}^q} \left[\frac{1}{n}\sum_{i=1}^n \log(1 + \gamma' g(z_i,\theta))\right]. \end{aligned} \tag{3.12}$$

The last expression again matches the characterization of the EL estimator provided in Section 2, showing that the somewhat mysterious "saddle point" formulation of the EL estimator provided there is a natural consequence of the Fenchel duality. According to the original definition, the EL estimator solves the two fold minimization problem (3.11); it is an estimator that minimizes a contrast function. But its dual form (3.12) replaces the second minimization in (3.11) with a (computationally more tractable) low-dimensional maximization problem, thereby yielding the saddle point formula (3.12).

Note also that the form of the contrast function corresponding to the choice $\phi(x) = -\log(x)$ is

$$D(\theta,\mu) = \inf_{P\in\mathcal{P}(\theta)} \int \log\frac{d\mu}{dP} d\mu = \inf_{P\in\mathcal{P}(\theta)} K(\mu, P),$$

where $K(P,Q) = \int \log \frac{dP}{dQ} dP$ denotes the Kullback-Leibler (KL) divergence between probability measures P and Q. That is, the EL estimator solves the minimization problem

$$\inf_{\theta\in\Theta} \inf_{P\in\mathcal{P}(\theta)} K(\mu_n, P) = \inf_{P\in\mathcal{P}} K(\mu_n, P). \tag{3.13}$$

The fact that empirical likelihood minimizes the KL divergence between the empirical measure μ_n and the moment condition model $\mathcal{P}$ plays an important role in the analysis of empirical likelihood with the large deviations theory presented in Section 4.

Choices of ϕ other than $-\log$ have been considered in the literature. Let $\phi(x) = x\log(x)$, then the contrast function evaluated at P is $D(P,\mu) = \int \log \frac{dP}{d\mu} dP = K(P,\mu)$. This is similar to empirical likelihood in that the contrast function is given by the KL divergence, but note that the roles of P and

μ are reversed. The Legendre transform of $\phi(x)$ is $\phi^\star(y) = e^{y-1}$. Using this in (3.9) and concentrating λ out, one obtains $\hat{\theta}$ as a solution to

$$\inf_{\theta\in\Theta}\max_{\gamma\in\mathbb{R}^q}\left[-\frac{1}{n}\sum_{i=1}^{n}e^{\gamma' g(z_i,\theta)}\right]. \tag{3.14}$$

This is the saddle-point estimator proposed by Kitamura and Stutzer (1997). It is sometimes called the exponential tilting estimator for θ. Note that $\phi'^{-1}(y) = e^{y-1}$ for this case, so the definition (3.10) yields

$$\widehat{\bar{P}(\hat{\theta})}(A) = \int_A e^{\hat{\gamma}' g(z_i,\hat{\theta})} d\mu_n \tag{3.15}$$

where $\hat{\gamma}$ is the parameter value at the saddle-point of (3.14).

Yet another popular choice of ϕ is $\phi(x) = \frac{1}{2}(x^2-1)$, which yields $D(P,\mu_n) = \frac{1}{2n}\sum_{i=1}^{n}(np_i-1)^2$. This is called the "Euclidean likelihood" by Owen (1991). Its Legendre transformation is $\phi^\star(y) = \frac{1}{2}(y^2+1)$. In this case the numerical optimization to evaluate the function $\widehat{v}^\star(\theta), \theta\in\Theta$ is unnecessary. Let $\bar{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n}g(z_i,\theta)$ and $\hat{S} = \frac{1}{n}\sum_{i=1}^{n}[g(z_i,\theta)-\bar{g}(\theta)][g(z_i,\theta)-\bar{g}(\theta)]'$. It is easy to see that for the quadratic $\phi^\star$ the maximization problem that defines $\widehat{v}^\star(\theta)$ (see (3.9)) has an explicit solution and the resulting GMC estimator solves

$$\inf_{\theta\in\Theta}\bar{g}(\theta)'\hat{S}^{-1}(\theta)\bar{g}(\theta).$$

Therefore, the choice $\phi(x) = \frac{1}{2}(x^2-1)$ leads to the continuous updating GMM estimator by Hansen, Heaton, and Yaron (1996); this connection between (continuous updating) GMM and Euclidean likelihood is noted in Kitamura (1996b).

Finally, Baggerly (1998), Kitamura (1996b) and Newey and Smith (2004) suggest using the Cressie-Read divergence family, which corresponds to the choice $\phi(x) = \frac{2}{\alpha(\alpha+1)}(x^{-\alpha}-1)$ indexed by the parameter α. The conjugate $\phi^\star$ of ϕ in this case is given by $\phi^\star(y) = -\frac{2}{\alpha}\left[-\frac{\alpha+1}{2}y\right]^{\frac{\alpha}{\alpha+1}} + \frac{2}{\alpha(\alpha+1)}$. Using this $\phi^\star$ in (3.9) yields the estimation procedure in Theorem 2.2 of Newey and Smith (2004). Parameter values $\alpha = -2, -1, 0$ and 1 yield Euclidean likelihood, exponential tilt, empirical likelihood and Pearson's χ^2, respectively.

3.2. **GMC and GEL.** It is unnecessary to use a specific function or a specific parametric family of functions for ϕ or $\phi^\star$ to define a GMC estimator $\hat{\theta}$. Kitamura (1996b) suggests a general family of estimator by considering ϕ's that are convex functions on $(0,+\infty)$. Alternatively, one may use the dual representation instead to define a class of estimators

$$\{\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta}\max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q}[\lambda - \frac{1}{n}\sum_{i=1}^{n}\phi^\star(\lambda+\gamma' g(z_i,\theta))] : \phi^\star \text{ is convex}\}.$$

If ϕ takes the form of the Cressie-Read family, then $\phi^\star$ is convex and homogeneous (plus an additive constant). Consequently, concentrating λ out and re-defining γ as γ/λ yields $\hat{\theta}$ as

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \max_{\gamma\in\mathbb{R}^q} \left[-\frac{1}{n}\sum_{i=1}^n \phi^\star(1+\gamma' g(z_i,\theta)) \right]. \tag{3.16}$$

Define $\rho(y) = -\phi^*(y+1)$, then

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \max_{\gamma\in\mathbb{R}^q} \left[\frac{1}{n}\sum_{i=1}^n \rho(\gamma' g(z_i,\theta)) \right]. \tag{3.17}$$

This is essentially equivalent to the Generalized Empirical Likelihood (GEL) estimator by Smith (1997), though his original derivation of GEL is based on an interesting application of the method of Chesher and Smith (1997). It is therefore quite different from the GMC-based derivation outlined above. Also, Smith's formulation of GEL demands only concavity on ρ in (3.17). The GEL family and the GMC family therefore do not completely coincide, though the difference between the two does not seem to matter much for practitioners as they both include commonly used estimators such as EL, exponential tilt and continuous updating as special cases. Smith (2004) provides a detailed account for GEL.

3.3. Some Properties.

The procedures based on GMC or GEL share some common properties. First, both family yield estimators that have the same asymptotic distribution as in (2.7) under reasonable conditions; see Kitamura and Stutzer (1997), Smith (1997), Imbens, Spady, and Johnson (1998) and Newey and Smith (2004). It is well known that the two step (optimal) GMM (Hansen (1982)) also yields the same first order asymptotics. Second, the value of the objective function can be used for inference. It has already been observed in Section 2 that one can construct a nonparametric analogue of the likelihood ratio statistic which has an appropriate χ^2-distribution asymptotically. This carries over to the procedures discussed in Section 3. For example, suppose one is interested in testing the null hypothesis that $\theta_0 \in \Theta_0 \subset \Theta, \dim(\Theta_0) = k - s$ in (2.2)[3], which puts s restrictions on the parameter space for θ. Under the hypothesis, the difference in the constrained and unconstrained objective function values obeys the following asymptotic distribution:

$$-2\left(\inf_{\theta\in\Theta} \widehat{v}(\theta) - \inf_{\theta\in\Theta_0} \widehat{v}(\theta) \right) \xrightarrow{d} \chi^2_s. \tag{3.18}$$

Third, a similar argument applies to overidentifying restrictions testing. One can use the maximum value of the GMC objective function to test the null hypothesis that $\mu \in \mathcal{P}$ (i.e. the model is correctly

[3] This requirement refers to the local dimension of Θ_0 at θ_0. Here and henceforth the symbol dim is often used to denote such local dimensions.

specified, or the overidentifying restrictions hold). Under this null,

$$\inf_{\theta \in \Theta} 2\widehat{v}(\theta) \xrightarrow{d} \chi^2_{q-k}.$$

Smith (2000) discusses various EL-based specification tests for (2.2). See also Ramalho and Smith (2002).

Asymptotic properties similar to those presented above also hold for the conventional two-step GMM (Hansen (1982)), but there are distinctive features of GMC/GEL that are not shared by the two-step GMM. Subsequent sections investigate those properties theoretically, though some informal arguments that have been often made in favor of GMC/GEL-type estimators are worth noting here.

The two-step GMM requires a preliminary estimator of the weighting matrix, which often causes problems in finite samples, whereas GMC or GEL avoids explicit estimation of it. Some theoretical advantages associated with this fact are discussed in Section 5, though one interesting consequence is the normalization invariance property of GMC/GEL. Suppose one obtains a moment condition of the form (2.2) as an implication of the economic model. It is obvious that one can replace g by $\tilde{g} = Ag$, where $A(\theta)$ is a nonsingular matrix that can depend on θ, and obtain an equivalent condition $E[\tilde{g}(z, \theta_0)] = 0, \theta_0 \in \Theta$. There should be no economic reason to prefer one representation over the other. The two-step GMM, however, yields different results in finite samples, depending on the choice of A. See Gali and Gertler (1999) for an example of this phenomenon in an actual empirical setting. All the estimators discussed in the previous section are invariant with respect to the choice of A.

The properties described above are interesting and desirable, though more theoretical developments are required to uncover decisive advantages of GMC/GEL estimators, and, in particular, those of empirical likelihood. This will be the main theme of the next two sections.

4. Large Deviations

One can pick an arbitrary convex function ϕ in GMC (or a concave function ρ in GEL) to define an estimator. This introduces a great deal of arbitrariness in estimating (2.2), and raises a natural and important question: which member of GMC (or GEL) should be used? Theoretical, practical and computational considerations are necessary to answer this question. This section and the next attempt to provide theoretical accounts, followed by more practical discussions in Section 8. Note that the results in this section provide a theoretical answer to the above question, but they have further implications. The optimality result here holds for a class of very general statistical procedures, including those which do not belong to GMC or GEL.

The conventional *first-order, local* asymptotic theory discussed in Sections 2 and 3 predicts identical asymptotic behavior for members of GMC/GEL estimators as far as the moment condition model (2.2) is correctly specified. Likewise, all comparable tests that belong to these families share identical properties under appropriate null and local alternative hypotheses. This is a consequence of the fundamental nature of the conventional asymptotic distribution theory, which relies on first order linear approximations. In reality, however, these estimators and tests can behave wildly differently in finite samples (see, for example, simulation results in Kitamura (2001) and Kitamura and Otsu (2005)). While the conventional asymptotic method is a useful device, it is important to explore approaches that go beyond local first order approximations to resolve these problems. At least two alternative approaches exist. One approach, taken by some researchers, is to explore local higher order asymptotic theory. This often yields useful and insightful results as will be discussed in Section 5, though it generally involves rather intricate calculations and delicate regularity conditions. Alternatively, first-order, global efficiency properties of GMC can be explored. Taking this approach enables us to evaluate the statistical implications of *global* and *nonlinear* structures of various GMC estimators not captured by local linear theory. This is a powerful tool for studying differences in the behavior of GMM, empirical likelihood and other estimators. Such an investigation belongs to the domain of the so-called large deviation theory, which is the theme of the current section.

The rest of this section covers two topics. The first is the large deviation theory of estimation in the moment condition model (2.2). The second is the large deviations analysis of various hypothesis testing problems, including inference concerning θ as well as testing the overidentifying restrictions of the model. Interestingly, in both cases empirical likelihood (i.e. GMC with $\phi(x) = -\log(x)$) yields optimality results, implying that empirical likelihood has a special status among competing procedures. Note, however, that obtaining an optimal estimator in terms of large deviations requires some modifications to the maximum empirical likelihood estimator discussed in Sections 2 and 3.

4.1. **Large Deviations and Minimax Estimation.** Consider again the moment condition model (2.2). θ_0, or its subvector, is the parameter of interest. The conventional asymptotic efficiency theory focuses on the behavior of estimators in a shrinking neighborhood of the true parameter value. In contrast, efficiency theory with LDP deals with a fixed neighborhood of the true value. For an estimator θ_n, consider the probability that it misses the true value θ_0 by a margin exceeding $c > 0$

$$\Pr\{\|\theta_n - \theta_0\| > c\}, \tag{4.1}$$

where $\|\cdot\|$ is the (Euclidean) norm.

The expression (4.1) can be interpreted as the expected risk $\mathrm{E}[L(\hat{\theta})]$ of the estimator θ_n under the loss function $L(\theta) = 1\{\|\theta_n - \theta_0\| > c\}$. It should be emphasized that other loss functions can be employed. It is generally possible to derive a large deviation optimal estimator under an alternative loss function, at least at the theoretical level. The treatment here focuses on the indicator loss function, however. It is a natural loss function, and commonly used in the literature (e.g. Bahadur (1964)). A nice feature of the indicator loss is that it leads to a practical and computationally convenient procedure, as discussed later in this section.

The parameter c in (4.1) is a loss function parameter that is chosen by the decision maker (i.e. the econometrician). In a typical empirical application, the econometrician would tolerate estimation errors within a certain margin. If one subscribes to the view that a model is an approximation of reality, it would be natural to allow a certain margin of error. Also, a number of authors argued the importance the concept of "economic significance" in econometrics; with that view, c can be chosen by considering a range within which errors are economically insignificant. In sum, c should be determined based on economic considerations. As an example, suppose the risk aversion parameter in a dynamic optimization model of consumers is being estimated. The econometrician then would have a range within which differences in the degree of risk aversion are economically not significant. The parameter c is a part of the econometrician's loss function and therefore should be decided based on the economic meaning of the parameter.

Once the parameter c is chosen, the next step is to make the probability (4.1) "small." The precise meaning of "small" will be defined shortly.

Evaluating (4.1) in finite samples is unrealistic unless the model is completely specified and extremely simple. On the other hand, simply letting n go to infinity is not informative, as the limit would be either 1 or 0 depending on whether θ_n is consistent or not. The theory of large deviations focuses on the asymptotic behavior of

$$(4.2) \qquad \left(\Pr\{\|\theta_n - \theta_0\| > c\}\right)^n .$$

or its logarithmic version

$$(4.3) \qquad \frac{1}{n} \log\left(\Pr\{\|\theta_n - \theta_0\| > c\}\right).$$

Letting n go to infinity in the latter gives the negative of the asymptotic decreasing rate of the probability that the estimator misses the true value by a margin that exceeds c. The goal would be then to make this limit as small as possible. The problem of minimizing the rate as in (4.3) has been considered in the context of parametric estimation, e.g. Bahadur (1960) and Fu (1973). These studies, however, usually require that the model belongs to the exponential family and do not extend to other models.

The moment condition model (2.2) is not a member of the exponential family, but there is a way to proceed. Kitamura and Otsu (2005) note that an asymptotic minimax criterion leads to an estimator for θ that possesses optimal properties in terms of large deviations, by using an approach proposed by Puhalskii and Spokoiny (1998). Consider, instead of (4.3), its maximum over all possible combinations of $(\theta,\, P)$:

$$\sup_{\theta\in\Theta}\sup_{P\in\mathcal{P}(\theta)}\frac{1}{n}\log\left(P^{\otimes n}\{\|\theta_n-\theta\|>c\}\right), \tag{4.4}$$

where $P^{\otimes n}=P\otimes P\otimes\cdots\otimes P$ denotes the n-fold product measure of P. (Since $z_i\sim_{iid}P$, the sample obeys the law $(z_1,...,z_n)\sim P^{\otimes n}$.)

Let $B\le 0$ denote an asymptotic lower bound for (4.4), that is,

$$\liminf_{n\to\infty}\inf_{\theta_n\in\mathcal{F}_n}\sup_{\theta\in\Theta}\sup_{P\in\mathcal{P}(\theta)}\frac{1}{n}\log\left(P^{\otimes n}\{\|\theta_n-\theta\|>c\}\right)\ge B, \tag{4.5}$$

where $\mathcal{F}_n$ denotes the set of all Θ-valued measurable functions of data $(z_1,...,z_n)$, i.e. the set of all estimators. The following constant B^* satisfies (4.5):

$$B^*=\sup_{Q\in\mathbf{M}}\inf_{\theta^*\in\Theta}\sup_{\theta\in\Theta:\|\theta^*-\theta\|>c}\sup_{P\in\mathcal{P}(\theta)}-K(Q,P). \tag{4.6}$$

(Recall $\mathbf{M}$ denotes the set of all probability measures on $\mathbb{R}^p$.) Moreover, this bound B^* turns out to be tight, therefore called the asymptotic minimax bound. The qualification "asymptotic" refers to $\liminf_{n\to\infty}$, whereas the term "minimax" corresponds to the operation $\inf_{\theta_n\in\mathcal{F}_n}\sup_{\theta\in\Theta}\sup_{P\in\mathcal{P}(\theta)}$.

Indeed, the minimax bound (4.6) can be achieved by an estimator based on empirical likelihood function (Kitamura and Otsu (2005)). Let $\hat{\theta}_{\mathrm{ld}}$ denote the minimizer of the objective function

$$Q_n(\theta)=\sup_{\theta^*\in\Theta:\|\theta^*-\theta\|>c}\ell(\theta^*),$$

where $\ell(\cdot)$ is the log empirical likelihood function defined in (2.5). The estimator $\hat{\theta}$ is a minimax estimator in the large deviation sense, as it reaches the bound B^* asymptotically:

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\sup_{P\in\mathcal{P}(\theta)}\frac{1}{n}\log\left(P^{\otimes n}\{\|\hat{\theta}_{\mathrm{ld}}-\theta\|>c\}\right)=B^*.$$

See Kitamura and Otsu (2005) for a proof. Note that $\hat{\theta}_{\mathrm{ld}}$[1] generally differs from the empirical likelihood estimator $\hat{\theta}_{\mathrm{EL}}$, unless, say, the sample empirical likelihood function $\ell(\cdot)$ is symmetric around $\hat{\theta}_{\mathrm{EL}}$.

Practical implementation of $\hat{\theta}_{\mathrm{ld}}$ is straightforward, at least when $k = \dim(\Theta)$ is low, since the objective function Q_n is a rather simple function of the log empirical likelihood $\ell(\cdot)$, whose numerical evaluation is easy (see Sections 2, 3 and 8).

If the dimension of θ is high, it is also possible to focus on a low dimensional sub-vector of θ and obtain a large deviation minimax estimator for it, treating the rest as nuisance parameters. This is potentially useful in practice, since it is often the case that a small number of "key parameters" in a model are economically interesting. Even in a case where every element of θ is important, one may want to apply the following procedure to each component of θ to lessen computational burden. Wlog, let $\theta = (\theta^{1\prime}, \theta^{2\prime})'$, where $\theta^1 \in \Theta^1$. Suppose the researcher chooses the loss function $1\{\|\theta^1 - \theta^1_n\| > c\}$ to evaluate the performance of an estimator θ^1_n for θ^1. The corresponding maximum (log) risk function is given by

$$\sup_{\theta\in\Theta}\sup_{P\in\mathcal{P}(\theta)}\frac{1}{n}\log\left(P^{\otimes n}\{\|\theta^1_n - \theta^1\| > c\}\right). \tag{4.7}$$

The limit inferior of the above display is bounded below by

$$B_1^* = \sup_{Q\in\mathbf{M}}\inf_{\theta^{1*}\in\Theta^1}\sup_{\theta\in\Theta:\|\theta^{1*}-\theta^1\|>c}\sup_{P\in\mathcal{P}(\theta)} -K(Q,P).$$

Let $\hat{\theta}^1_{\mathrm{ld}}$ minimize the function

$$Q^1_n(\theta^1) = \sup_{\theta^*\in\Theta:\|\theta^{1*}-\theta_1\|>c}\ell(\theta^*).$$

This is a minimax estimator, in the sense that it achieves the lower bound of the maximum risk (4.7) asymptotically:

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\sup_{P\in\mathcal{P}(\theta)}\frac{1}{n}\log\left(P^{\otimes n}\{\|\hat{\theta}^1_{\mathrm{ld}} - \theta^1\| > c\}\right) = B_1^*.$$

If the parameter of interest θ_1 is scalar, it is possible to provide an interesting and practically useful characterization of the minimax estimator $\hat{\theta}^1_{\mathrm{ld}}$. Assume, for the sake of argument, that the minimum of Q^1_n is attained uniquely by $\hat{\theta}^1_{\mathrm{ld}}$. Then

$$\begin{aligned}\hat{\theta}^1_{\mathrm{ld}} &= \operatorname*{argmin}_{\theta^1\in\Theta^1}\sup_{\theta^*\in\Theta:|\theta^{1*}-\theta^1|>c}\ell(\theta) \\ &= \operatorname*{argmin}_{\theta^1\in\Theta^1}\sup_{\theta_1^*\in\Theta^1:|\theta^{1*}-\theta^1|>c}\sup_{\theta^{2*}}\ell(\theta^{1*},\theta^{2*}) \\ &= \operatorname*{argmin}_{\theta^1\in\Theta^1}\sup_{\theta_1^*\in\Theta^1:|\theta^{1*}-\theta^1|>c}\ell_1(\theta^{1*}),\end{aligned} \tag{4.8}$$

where $\ell^1(\theta^1) = \sup_{\theta^2} \ell(\theta^1, \theta^2)$ is the log empirical likelihood function with the nuisance parameter θ^2 profiled out. Imagine the function $\ell^1(\cdot)$ plotted against the parameter space Θ^1 of θ^1, which is (a subset of) $\mathbb{R}$. Choose a level set of $\ell^1(\cdot)$ so that its length is $2c$, then the estimator $\hat{\theta}^1_{\text{ld}}$ is the midpoint of the level set. To see this, notice that the last expression in (4.8) indicates that the value of $\hat{\theta}^1_{\text{ld}}$ is chosen so that the maximum of ℓ^1 outside of the c-ball with center $\hat{\theta}^1_{\text{ld}}$ becomes as small as possible. If an alternative value $\check{\theta}^1$ in place of $\hat{\theta}^1_{\text{ld}}$ is used, the c-ball around it will exclude some points where the values of $\ell^1(\cdot)$ are higher than $Q^1_n(\hat{\theta}^1_{\text{ld}})$, making $Q^1_n(\check{\theta}^1)$ larger than $Q^1_n(\hat{\theta}^1_{\text{ld}})$. This is true for any $\check{\theta}_1 \neq \hat{\theta}^1_{\text{ld}}$, therefore $\hat{\theta}^1_{\text{ld}}$ minimizes Q^1_n.

The above characterization of $\hat{\theta}^1_{\text{ld}}$ implies that the estimator can be interpreted as a "robustified" version of the original empirical likelihood estimator. Again, imagine the profiled empirical likelihood function ℓ^1 plotted against the space of θ^1. Suppose the function has a "plateau" of length $2c$. This should include the maximizer of ℓ^1, but the maximum can occur at a point that is close to one of the end points of the plateau. The empirical likelihood estimator "follows" small fluctuations over the plateau, since it has to correspond to the exact maximum. In contrast, the minimax estimator always chooses the center of the plateau and therefore is robust against these small fluctuations. This is reminiscent of arguments that favor a posterior mean Bayes estimator over MLE, on the ground that the former takes a weighted average of the likelihood function and is more robust against sample fluctuations of the likelihood function than the latter. This interpretation of $\hat{\theta}^1_{\text{ld}}$ as a robustified estimator applies to the case where θ^1 is multi-dimensional as well.

The preceding discussion described the new procedure as a point estimator, though it may be better understood as a fixed-length interval estimation method. The concept of fixed-length interval estimators can be found, for example in Wald (1950). As before, let θ^1 denote the parameter of interest in the vector θ and suppose it is a scalar. Consider the set $\mathcal{I}_n$ of interval estimators I_n of length $2c$ for θ^1. From the above discussion, the large deviation probability of such an interval not containing the true value is asymptotically bounded from below by B^*_1, which is attained by the interval estimator $\hat{I}_{\text{ld}} = [\hat{\theta}^1_{\text{ld}} - c, \hat{\theta}^1_{\text{ld}} + c]$. That is,

$$\liminf_{n\to\infty} \inf_{I_n \in \mathcal{I}_n} \sup_{\theta\in\Theta} \sup_{P\in\mathcal{P}(\theta)} \frac{1}{n} \log\left(P^{\otimes n}\{\theta^1 \notin I_n\}\right) = B^*_1 = \lim_{n\to\infty} \sup_{\theta\in\Theta} \sup_{P\in\mathcal{P}(\theta)} \frac{1}{n} \log\left(P^{\otimes n}\{\theta^1 \notin \hat{I}_{\text{ld}}\}\right).$$

The interval estimator $\hat{I}_{\text{ld}}$ is therefore optimal. $\hat{\theta}^1_{\text{ld}}$ is not necessarily consistent when viewed as a point estimator, but the corresponding $\hat{I}_{\text{ld}}$ is consistent in the sense that it contains the true parameter value with probability approaching 1.

The choice of the parameter c should be determined by the goal of the economic analysis as discussed at the beginning of this section, though some further remarks on this issue are in order.

First, experimental results from Kitamura and Otsu (2005) suggest that a wide range of values of c work for realistic sample sizes. See Section 8 for more information on finite sample properties of the minimax estimator and other estimators.

Second, it is reasonable to assume that the researcher would choose a smaller value of c, when a larger data set is available and therefore more accurate estimation would be possible. If one calculates $\hat{\theta}^1_{\mathrm{ld}}$ for a sequence of constants $\{c_n\}_{n=1}^{\infty}$ that converges to 0 slowly, the estimator is consistent as a point estimator, while it may be still possible to show that it has an asymptotic optimality property. Such an investigation would involve the theory of moderate deviations, which has been applied to estimation problems; see, for example, Kallenberg (1983).

4.2. **Minimax Testing.** Section 4.1 applied an asymptotic minimax approach to parameter estimation. Kitamura and Otsu (2005) show that the similar approach (cf. Puhalskii and Spokoiny (1998)) leads to a testing procedure that has a large deviation minimax optimality property in the model (2.2). Let Θ_0 be a subset of the parameter space Θ of θ. Consider testing

$$\mathrm{H}_0 : \theta \in \Theta_0$$

against

$$\mathrm{H}_1 : \theta \in \Theta_0^c$$

(Θ_0^c denotes the complement of Θ_0 in Θ). To derive a minimax test, a few decision theoretic concepts are useful. The econometrician observes the data $\{z_1, ..., z_n\}$ to reach a decision to accept H_0 or reject it in favor of H_1. So it can be represented by a binary-valued function $d_n = d_n(z_1, ..., z_n)$, taking the value of 0 if H_0 is accepted and the value of 1 otherwise. An appropriate loss function for decision d_n is

$$L(d_n) = w1\{d_n = 1, \mathrm{H}_0 \text{ holds}\} + (1 - w)1\{d_n = 0, \mathrm{H}_1 \text{ holds}\}$$

where the weighting factor w belongs to $[0, 1]$. The econometrician chooses w; as seen below, this parameter determines the critical value of the minimax test. Applying the same argument as in Section 4.1 yields a decision function that is large deviation minimax optimal. Let $\mathcal{P}_0 = \cup_{\theta\in\Theta_0}\mathcal{P}(\theta)$ and $\mathcal{P}_1 = \cup_{\theta\in\Theta_1}\mathcal{P}(\theta)$. The maximum log expected risk, normalized by $\frac{1}{n}$ is:

$$\frac{1}{n}\max\left\{\log(w^n \sup_{P\in\mathcal{P}_0} P^{\otimes n}\{d_n = 1\}), \log((1-w)^n \sup_{P\in\mathcal{P}_1} P^{\otimes n}\{d_n = 0\})\right\}.$$

This corresponds to (4.4) in the estimation problem. The limit inferior of the above display (as n tends to infinity) is bounded below by

$$C^* = \sup_{Q \in \mathbf{M}} \min \left\{ \log w + \sup_{P \in \mathcal{P}_0} -K(Q,P), \log(1-w) + \sup_{P \in \mathcal{P}_1} -K(Q,P) \right\}.$$

It can be shown that this large deviation bound is attained by the following decision function

$$\hat{d}_{\mathrm{ld}} = 1\{\log \frac{w}{1-w} < \sup_{\theta \in \Theta} l(\theta) - \sup_{\theta \in \Theta_0} l(\theta)\}, \tag{4.9}$$

that is,

$$\lim_{n \to \infty} \frac{1}{n} \max \left\{ \log(w^n \sup_{P \in \mathcal{P}_0} P^{\otimes n}\{\hat{d}_{\mathrm{ld}} = 1\}), \log((1-w)^n \sup_{P \in \mathcal{P}_1} P^{\otimes n}\{\hat{d}_{\mathrm{ld}} = 0\}) \right\} = C^*$$

if $\frac{1}{2} \le w \le 1$. (If the econometrician chooses w that is less than $\frac{1}{2}$, $\hat{d}_{\mathrm{ld}}$ needs to be modified so that the first supremum in the above definition of $\hat{d}_{\mathrm{ld}}$ is taken over Θ_1.) The testing procedure (4.9) is the empirical likelihood ratio (ELR) test for H_0 with critical value $2 \log \frac{w}{1-w}$ as described in Section 2; see also Equation (3.18). That is, ELR is minimax, large deviation optimal for testing H_0 against H_1.

4.3. **GNP-Optimal Testing.** Specification analysis of moment condition models is often carried out using the GMM-based overidentifying restrictions test by Hansen (1982). The null hypothesis of Hansen's test takes the form (2.2), that is, the moment condition holds for *some* θ in Θ. It can be expressed using the notation in Section 3.3 as:

$$\mu \in \mathcal{P}. \tag{$\mathbf{H}$}$$

Previous sections introduced alternative tests for (**H**), including the ELR overidentification test based on $\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})$; see Equation (2.9) and discussions thereafter.

Again, the existence of alternative procedures raises the question of which test should be used. Various asymptotic efficiency criteria can be applied for comparing competing tests for (**H**). See Serfling (1980) for a comprehensive catalog of asymptotic efficiency criteria. Among them, the well-known Pitman efficiency criterion uses local first-order approximations and is not informative here, for the same reason discussed at the beginning of this section. The asymptotic efficiency criterion by Hoeffding (1963), however, reveals that the ELR test is optimal in a asymptotic large deviations sense. This conclusion, obtained by Kitamura (2001), is in a sense stronger than the results in the previous section on parametric hypothesis testing. It claims that the ELR is *uniformly* most powerful, whereas the previous result is about minimax optimality. This kind of property is sometimes called a Generalized Neyman-Pearson (GNP) type lemma.

As before, a test is represented by a sequence of binary functions $d_n = d_n(z_1, ..., z_n), n = 1, 2, ...$ which takes the value of 0 if the test accept (**H**) and 1 otherwise. The conventional asymptotic power comparison is based on the type II error probabilities of tests that have comparable type I error probabilities. Hoeffding (1963) takes this approach as well, but he evaluates type I and type II errors using LDP. In the present context, size properties of competing tests are made comparable by requiring that, for a parameter $\eta > 0$, each test d_n satisfies

$$\textbf{(L)} \qquad \sup_{P \in \mathcal{P}} \limsup_{n \to \infty} \frac{1}{n} \log P^{\otimes n}\{d_n = 1\} \leq -\eta$$

Therefore η determines the level of a test via large deviations.

Now, use the η in (**L**) to define the ELR test as follows:

$$d_{\mathrm{ELR},n} = \begin{cases} 0 & \text{if } \frac{1}{2n}\mathrm{elr}(\hat{\theta}_{\mathrm{EL}}) \leq -\eta \\ 1 & \text{otherwise.} \end{cases}$$

Kitamura (2001) shows the following two facts under weak regularity conditions.

(I) $d_{\mathrm{ELR},n}$ satisfies the condition (**L**).

(II) For every test d_n that satisfies (**L**),

$$\limsup_{n \to \infty} \frac{1}{n} \log P^{\otimes n}\{d_n = 0\} \geq \limsup_{n \to \infty} \frac{1}{n} \log P^{\otimes n}\{d_{ELR,n} = 0\}$$

for every $P \notin \mathcal{P}$.

Fact (I) shows that the large deviation rate of the type I error probability of the ELR test defined as above satisfies the size requirement (**L**). The left hand side and the right hand side of the inequality in Fact (II) correspond to the LDP of the type II errors of the arbitrary test d_n and that of the ELR test $d_{\mathrm{ELR},n}$, respectively. The two facts therefore mean that, among all the tests that satisfies the LDP level condition (**L**) (and the regularity conditions discussed in Kitamura (2001)), there exists no test that outperforms the ELR test in terms of the large deviation power property. Note that Fact (II) holds for every $P \notin \mathcal{P}$, therefore ELR is uniformly most powerful in terms of LDP. That is, ELR is a GNP test.

To illustrate this result, take the simplest example: suppose $z_i \sim_{iid} P, i = 1, ..., n$ and $E[z_i] = m \in \mathbb{R}$. The null hypothesis $m = m_0$ is tested against $m \neq m_0$. In absence of further distributional assumptions, a standard procedure described in textbooks is to carry out a large sample test based on the statistic $n(\bar{z} - m_0)^2/\hat{s}_z$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ and $\hat{s}_z$ is a consistent estimator for the variance of z. The above result shows that the standard procedure as above is suboptimal, since the ELR is the uniformly most powerful in Hoeffding's criterion. The simulation experiments reported in Section 8.2.2 considers this setting and provides strong support for the theoretical implications described above.

4.4. **EL, Large Deviations, and Sanov's Theorem.** The analysis in the previous two sections shows various efficiency properties of empirical likelihood in terms of large deviations. This phenomenon is by no means a coincidence. The fundamental reason why empirical likelihood emerges as an optimal procedure comes from a LDP for empirical measures, called Sanov's theorem (Sanov (1961); see also Deuschel and Stroock (1989), Theorem 3.1.17.). Suppose $z_i \sim_{iid} \mu, i = 1, ..., n$, and equip the space of all probability measures $\mathbf{M}$ with the topology of weak convergence. For an arbitrary set $\mathcal{G} \in \mathbf{M}$, let $\mathcal{G}^o$ and $\bar{\mathcal{G}}$ denote the interior and the closure of $\mathcal{G}$, respectively. Sanov's Theorem shows that the empirical measure μ_n satisfies

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr(\mu_n \in \mathcal{G}^o) \geq - \inf_{\nu \in \mathcal{G}^o} K(\nu, \mu)$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr(\mu_n \in \bar{\mathcal{G}}) \leq - \inf_{\nu \in \bar{\mathcal{G}}} K(\nu, \mu).$$

Put loosely, the probability that the empirical measure falls into the set $\mathcal{G}$ is governed by the minimum value of the Kullback-Leibler divergence number between the probability measure and $\mathcal{G}$. The moment condition model is represented by the set $\mathcal{P}$, so it is reasonable to expect that using the minimum KL divergence $\inf_{P\in\mathcal{P}} K(\mu, P)$, or more precisely, its empirical version $\inf_{P\in\mathcal{P}} K(\mu_n, P)$, as a statistical criterion leads to optimal procedures. As seen in (3.13), however, empirical likelihood solves the empirical KL minimization problem and therefore often achieves optimality in a large deviations sense. The choice of $\phi(x) = -\log(x)$ emerges naturally from the LDP, not as a consequence of an arbitrary and artificial choice of an econometric objective function.

5. Higher Order Theory

The LDP-based approach, presented in the previous section, utilized global first order approximations of empirical likelihood and other related methods. This section presents some results from an alternative approach based on local higher order approximations. Interestingly, the two quite different approaches tend to yield similar conclusions; empirical likelihood often exhibits desirable properties in terms of higher order comparisons as well, though there are some differences in the conclusions obtained from the two approaches.

5.1. **Estimation.** Newey and Smith (2004) investigate higher order properties of the GEL family of estimators. They find that GEL estimators have good properties in terms of the second order bias.

Moreover, empirical likelihood has a special bias reducing property among the GEL estimators. To illustrate their findings, it is instructive to consider the conventional two step GMM estimator for (2.2) and compare it with GEL estimators. Write $\bar{D}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g(z_i, \theta)$, $\bar{S}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta) g(z_i, \theta)'$ and $\bar{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta)$. The two step GMM, denoted by $\hat{\theta}_{\mathrm{GMM}}$, based on a preliminary estimator $\check{\theta}$ is a root of the following first order condition:

$$\bar{D}(\hat{\theta}_{\mathrm{GMM}})'\bar{S}(\check{\theta})^{-1}\bar{g}(\hat{\theta}_{\mathrm{GMM}}) = 0. \tag{5.1}$$

This can be regarded as a feasible version of the infeasible optimally weighted sample moment condition with $D = E[\nabla_{\theta_0} g(z, \theta)], S = E[g(z_i, \theta_0)g(z_i, \theta_0)']$, which would yield an "ideal" estimator $\hat{\theta}_{\mathrm{ideal}}$ as its root:

$$D'S^{-1}\bar{g}(\hat{\theta}_{\mathrm{ideal}}) = 0. \tag{5.2}$$

The effect of replacing D and S with $\bar{D}(\hat{\theta}_{\mathrm{GMM}})$ and $\bar{S}(\check{\theta})$ is negligible in the conventional first order asymptotics by Slutsky's theorem. These terms, however, do affect the bias of $\hat{\theta}_{\mathrm{GMM}}$ of order $O(\frac{1}{n})$ for two reasons. First, even evaluated at the true value θ_0, these two sample moments are correlated with the sample mean of g, and these correlations show up in the second order bias term of $\hat{\theta}_{\mathrm{GMM}}$. Second, the effect of the preliminary estimator $\check{\theta}$ also appears in the second order bias term. In particular, the first effect from the correlations tend to grow with the number of moment conditions q. See Newey and Smith (2004), Donald, Imbens, and Newey (2003) and Imbens and Spady (2006).

The situation changes for the EL estimator $\hat{\theta}_{\mathrm{EL}}$. Appendix shows that the first order condition for $\hat{\theta}_{\mathrm{EL}}$, using the notation $\hat{D}(\theta) = \sum_{i=1}^{n} \hat{p}_{\mathrm{EL}i} \nabla_\theta g(z_i, \theta)$ and $\hat{S}(\theta) = \sum_{i=1}^{n} \hat{p}_{\mathrm{EL}i} g(z_i, \theta)g(z_i, \theta)'$, can be written as:

$$\hat{D}(\hat{\theta}_{\mathrm{EL}})'\hat{S}^{-1}(\hat{\theta}_{\mathrm{EL}})\bar{g}(\hat{\theta}_{\mathrm{EL}}) = 0; \tag{5.3}$$

see Theorem 2.3 of Newey and Smith (2004) as well as Donald and Newey (2000). This is similar to the first order condition (5.1) for GMM, though there are important differences. Notice that D and S that appear in the "ideal" first order condition (5.2) are estimated by $\hat{D}(\hat{\theta}_{\mathrm{EL}})$ and $\hat{S}(\hat{\theta}_{\mathrm{EL}})$ in (5.3). These are semiparametrically efficient estimators of D and S under the moment restriction (2.2), as discussions in Section 2 imply. This means that they are asymptotically uncorrelated with $\bar{g}(\theta_0)$, removing the important source of the second order bias of GMM. Moreover, the EL estimator does not involve a preliminary estimator, thereby eliminating the other source of the second order bias in GMM mentioned above. Newey and Smith (2004) formalize this intuition and obtain an important conclusion that the second order bias of the EL estimator is equal to that of the infeasible "ideal" estimator $\hat{\theta}_{\text{ideal}}$. Schennach (2004) and Ragusa (2005) present interesting higher-order asymptotic results that are closely related to those of Newey and Smith (2004).

Some, if not all, of the nice second bias properties of EL are shared by other members of the GEL (or GMC) family. Newey and Smith (2004) observe that the Jacobian term D in the first order condition of GEL estimators is efficiently estimated, therefore the second order bias term due to the correlation between $\bar{g}$ and the estimator for D is absent. Also, they are free from second order bias from preliminary estimation, because they are one-step estimators. Therefore they possess merits over GMM in terms of higher order bias due to these factors.

In general, however, members of GEL other than EL have first order conditions where S is not efficiently estimated, and this can potentially cause bias through its correlation with $\bar{g}$. To see this point, take the continuous updating GMM estimator (CUE), which is a member of GMC as discussed in Section 3. Let $\tilde{D}(\theta) = \nabla_\theta \bar{g}(\theta) - \left(\frac{1}{n}\sum_{i=1}^n \nabla_\theta g(z_i,\theta) g(z_i,\theta)\right) \bar{S}^{-1}(\theta)\bar{g}(\theta)$, then the first order condition for CUE is

$$\tilde{D}(\hat{\theta}_{\text{cue}})'\bar{S}^{-1}(\hat{\theta}_{\text{cue}})\bar{g}(\hat{\theta}_{\text{cue}}) = 0 \tag{5.4}$$

(see Appendix). The term subtracted from $\nabla_\theta \bar{g}(\hat{\theta}_{\text{cue}})$ in $\tilde{D}(\hat{\theta}_{\text{cue}})$ makes it a semiparametrically efficient estimator for D. $\bar{S}(\hat{\theta}_{\text{cue}})$, however, is just a sample average and therefore not an efficient estimator for S; see Brown and Newey (2002). This, in turn, contributes to the second order bias of the continuous updating GMM. It has been recognized that the effect of the *estimated* weighting matrix of GMM is an important source of bias; see Altonji and Segal (1996) for an experimental study on this problem. Altonji and Segal (1996), based on their finding, recommend using the $q-$dimensional identify matrix for weighting $\bar{g}$ to avoid this problem, though this solution has the cost of not being efficient asymptotically. The higher order theory reviewed in this section suggests that empirical likelihood successfully addresses this problem, whereas other GEL or GMC estimators in general do not. Newey and Smith (2004) also analyze higher order MSE of bias-corrected GMM and GEL and note that the bias-corrected empirical likelihood estimator is third-order efficient. A recent paper by Kunitomo and Matsushita (2003) provides a detailed numerical study of EL and GMM, emphasizing on cases where the number of moments q is large. They find that the distribution of the EL estimator tends to be more centered and concentrated around the true parameter value compared with that of GMM. They also report that the asymptotic normal approximation appears to be more appropriate for EL than for GMM.

In a related recent study, Newey and Windmeijer (2006) consider asymptotics of GEL and GMM under "many weak moment conditions." An interesting aspect of this asymptotic scheme is that it captures an additional variance term in the asymptotic distribution of GEL due to the randomness in the (implicitly) estimated Jacobian term D. They find that the two-step GMM is asymptotically biased under this scheme, whereas GEL is not. They further propose an appropriate variance estimator for GEL in this case.

5.2. **Testing.** One of the significant findings in the early literature of empirical likelihood is the Bartlett correctability of the empirical likelihood ratio test, discovered by DiCiccio, Hall, and Romano (1991). A well-known result for parametric likelihood shows that one can improve the accuracy of the parametric likelihood ratio (LR) test by adjusting it by a constant called the Bartlett factor. DiCiccio, Hall, and Romano (1991) prove that this result holds for empirical likelihood. They consider testing a hypothesis of the form $\theta = \theta_0$ in (2.2) when the model is just-identified, i.e. $q = k$. The relationship (2.10) implies that the empirical likelihood ratio statistic r for the constraint and $\mathrm{elr}(\theta_0)$ are identical for this case. Recall that the asymptotic distribution of $\mathrm{elr}(\theta_0)$ (see (2.9)) is chi-square with q degrees of freedom, i.e. $\Pr\{\mathrm{elr}(\theta_0) \le x\} \to \Pr\{\chi^2_q \le x\}$ for $x \ge 0$ as $n \to \infty$. It can be shown that the accuracy of this approximation is of order n^{-1}:

$$\Pr\{\mathrm{elr}(\theta_0) \le x\} = \Pr\{\chi^2_q \le x\} + O(n^{-1}).$$

The error rate n^{-1} is good but not surprising since it can be achieved by other conventional tests, such as the Wald test. What is surprising about the Bartlett correctability result discovered by DiCiccio, Hall, and Romano (1991) is that the ELR test, which is nonparametric, permits Bartlett correction and it yields the same accuracy rate as in the parametric case. Let a denote the Bartlett factor, then

$$\Pr\{\mathrm{elr}(\theta_0)(1 + n^{-1}a) \le x\} = \Pr\{\chi^2_q \le x\} + O(n^{-2}).$$

See DiCiccio, Hall, and Romano (1991) for an analytical expression of a. The Bartlett factor can be replaced by an appropriate estimator without affecting the error rate n^{-2}. Notice that no element of the parameter vector θ is estimated in the testing problem considered by DiCiccio, Hall, and Romano (1991). Showing Bartlett correctability of ELR for more complex testing problems is harder, though some progress has been made. For example, Chen and Cui (2005) consider Bartlett correctability in the case where the model is just identified ($q = k$), and one is interested in testing an s-dimensional subvector of the parameter θ with $s < k$. See also Chen and Cui (2004). Though their result is stated in terms of just identified models, the Chen-Cui theorem immediately implies the Bartlett correctability of the empirical likelihood-based overidentifying restrictions test statistic $\text{elr}(\hat{\theta}_{\text{EL}})$, which is important in econometric applications. Whang (2006) reports strongly supportive experimental evidence for the Bartlett correction for quantile regression models. See also Chen, Leung, and Qin (2003) for simulation results on Bartlett correction in an interesting application of EL. On the other hand, it should be added that Corcoran, Davison, and Spady (1995) raise a question concerning the empirical relevance of Bartlett correction. These issues deserve further investigation.

6. Some variations of EL

6.1. **Estimation under Conditional Moment Restrictions.**

The moment condition model (2.2) is standard, though it is sometimes useful to consider a model stated in terms of a conditional moment restriction. Suppose, instead of (2.2), random variables x and z satisfies the condition

$$E[g(z,\theta)|x] = 0, \theta \in \Theta. \tag{6.1}$$

This is trivially satisfied for the standard mean regression model $E[y|x] = m(x,\theta)$ by setting $g(z,\theta) = y - m(x,\theta)$ and $z = (x,y)$. It also holds for many models of dynamic optimization, where (6.1) is interpreted as a stochastic Euler equation. The condition (6.1) implies (2.2), thus the former is stronger than the latter. This feature often leads to a common practice where a researcher picks an arbitrary matrix-valued function $a(x)$ of x as a matrix of instruments, then applies GMM, EL or other methods to an implication of (2.2):

$$E[a(x)g(z,\theta)] = 0.$$

Such a procedure is used under the presumption that the chosen instrument $a(x)$ identifies θ, which is not necessarily true even if θ is identified in the original model (6.1); see Dominguez and Lobato (2004) on this issue and other identification problems in the standard treatment of conditional moment restriction models. Moreover, it fails to fully utilize the information contained in the conditional moment restriction, and the resulting estimator does not achieve the semiparametric efficiency bound in general. A more satisfactory approach is to directly impose (6.1) in estimating θ.

Let $D(x) = E[\nabla_\theta g(z,\theta)]$ and $V(x) = E[g(z,\theta)g(z,\theta)'|x]$. Chamberlain (1987) shows that the semiparametric efficiency bound for the model (6.1) is given by

$$\mathcal{I}^{-1} = \left[E[D(x)'V^{-1}(x)D(x)]\right]^{-1}, \tag{6.2}$$

which can be attained by setting $a^*(x) = D'(x)V^{-1}(x)$ as instruments. One way to achieve this bound in practice is to apply a two step procedure. In the first step, one obtains an inefficient preliminary estimator $\tilde{\theta}$ for θ, and the unknown functions $D(x)$ and $V(x)$ are estimated by running nonparametric regressions of $\nabla_\theta g(z,\tilde{\theta})$ and $g(z,\tilde{\theta})g'(z,\tilde{\theta})$ on x. The nonparametric estimates $\tilde{D}(x)$ and $\tilde{V}(x)$ evaluated at $x_i, i = 1,..,n$ are used to construct estimated optimal instruments $\tilde{a}^*(x_i), i = 1,...,n$. In the second step, the optimal GMM is implemented using $\tilde{a}^*(x_i), i = 1,...,n$ as instruments. See Robinson (1987) and Newey (1990) for details of this approach.

The two-step approach is asymptotically valid under relatively mild conditions, but it is important to explore alternative approaches based on empirical likelihood for the following reasons. First, the validity of the two-step approach relies on the availability of a preliminary consistent estimator for θ. This can be in principle achieved by choosing an appropriate function $a(x)$ that identifies θ, but this is by no means guaranteed, as noted in the study by Dominguez and Lobato (2004) mentioned above. Second, Dominguez and Lobato (2004) also find that even if the form of the optimal IV a^* were known, the resulting moment condition $E[a^*(x)g(z,\theta)] = 0$ may fail to identify θ while the original model (6.1) identifies it. Third, the theoretical analysis in previous sections on unconditional moment condition models shows that GMC/GEL-type methods — empirical likelihood in particular — have various advantages over two-step procedures.

The rest of this subsection discusses two distinct empirical likelihood-based approaches to conditional moment restriction models, one proposed by Kitamura, Tripathi, and Ahn (2004) and the other by Donald, Imbens, and Newey (2003). The first uses kernel smoothing or a similar nonparametric regression technique to incorporate local restrictions implied by (6.1). The second employs an expanding set of unconditional moment restrictions so that the conditional moment restriction is "spanned" asymptotically. Both EL-based approaches address the two issues for the two-step approach regarding identification due to the choice of instruments; see more discussions below.

The approach taken by Kitamura, Tripathi, and Ahn (2004) utilizes kernel regression to calculate localized empirical log-likelihood, though other nonparametric regression techniques work for the purpose as well. Let K and h be an appropriate kernel function and bandwidth, and define a version of nonparametric log-likelihood localized at x_i,

$$(6.3) \qquad \ell_{\mathrm{LNP}i}(p_{i1},...,p_{in}) = \sum_{j=1}^{n} w_{ij}\log p_{ij}, \quad w_{ij} = \frac{K(\frac{x_i - x_j}{h})}{\sum_{j=1}^{n} K(\frac{x_i - x_j}{h})}, \quad (p_{i1},...,p_{in}) \in \Delta,$$

which is to be maximized subject to the conditional mean zero constraint for a given value of θ:

$$\sum_{j=1}^{n} p_{ij} g(z_j,\theta) = 0.$$

The idea of applying a nonparametric regression technique to likelihood is reminiscent of the expected log likelihood criterion that justifies the local likelihood methodology (Hastie and Tibshirani (1986)). The maximum value of the above optimization problem is used as the empirical log-likelihood contribution of the $i-$th observation, and in what follows denoted by $\ell_i(\theta)$. The duality result in Section 3 shows that

$$\ell_i(\theta) = w_{ij} \log w_{ij} - \max_{\gamma_i \in \mathbb{R}^q} \sum_{j=1}^{n} w_{ij} \log(1 + \lambda_i' g(z_j, \theta)). \tag{6.4}$$

The dual form formulation (6.4) is obviously preferred over the primal form formulation (6.3) from the computational point of view. Define the log-likelihood function for $\theta \in \Theta$ conditional on $X = \{x_i\}_{i=1}^n$ as

$$\ell_{\mathrm{CEL}}(\theta) = \sum_{i=1}^{n} \ell_i(\theta). \tag{6.5}$$

The maximizer of $\ell_{\mathrm{CEL}}(\theta)$ may be termed the conditional empirical likelihood estimator (or the smoothed empirical likelihood estimator) for θ for obvious reasons,[4] and will be denoted by $\hat{\theta}_{\mathrm{CEL}}$. See also LeBlanc and Crowley (1995) and Zhang and Gijbels (2003) for similar estimators.

Kitamura, Tripathi, and Ahn (2004) show that the limiting distribution of the conditional empirical likelihood estimator $\hat{\theta}_{\mathrm{CEL}}$ is given by

$$\sqrt{n}(\hat{\theta}_{\mathrm{CEL}} - \theta) \xrightarrow{d} \mathrm{N}(0, \mathcal{I}^{-1}),$$

that is, it achieves the semiparametric efficiency bound defined by (6.2). Unlike a two stage procedure where the choice of instruments in the first stage can affect identifiability, this estimator directly exploits the identification power of (6.1). Even in the examples presented by Dominguez and Lobato (2004) where the optimal IV a^* fails to deliver identification, the conditional empirical likelihood estimator is consistent as far as the original model (6.1) identifies θ. One may find this fact rather paradoxical, though this is due to the global properties of the objective functions. The likelihood nature of ℓ_{CEL} guarantees that its value is globally maximized at the true value asymptotically. In contrast, the form of optimal IV is based on local efficiency considerations, therefore the objective function of the corresponding GMM with optimal IV may fail to deliver identification. Also, the estimator $\hat{\theta}_{\mathrm{CEL}}$ avoids explicit estimation of the functions $D(x)$ and $V(x)$. This feature also applies

[4] Kitamura, Tripathi, and Ahn (2004) allow the support of x to be unbounded, and they use trimming to deal with technical problems associated with it. The treatment in this section ignores this issue to simplify presentation. Kitamura, Tripathi, and Ahn (2004) report that the use of trimming factors did not affect the result of their simulation experiments qualitatively.

to inference, i.e. testing and confidence interval calculation. To test a hypothesis of the form $\theta \in \Theta_0, \dim(\Theta_0) = k - s$, form a likelihood ratio statistic based on ℓ_{CEL}:

$$r_{\text{CEL}} = -2 \left(\sup_{\theta \in \Theta_0} \ell_{\text{CEL}}(\theta) - \sup_{\theta \in \Theta} \ell_{\text{CEL}}(\theta) \right).$$

This converges to a χ^2_s random variable in distribution under the null hypothesis. The same result, of course, can be used for constructing confidence intervals by inverting the likelihood ratio statistic.

Kitamura, Tripathi, and Ahn (2004) report some Monte Carlo results of this estimator and existing two-step estimators. The conditional EL estimator $\hat{\theta}_{\text{CEL}}$ performs remarkably well, and often works substantially better than the two step estimators in their simulations. For example, the precision of the conditional EL estimator, in terms of various dispersion measures, is close to that of the infeasible ideal estimator based on the unknown optimal IV even for a moderate sample size. They also report that the likelihood-ratio test based on r_{CEL} works well, in terms of its size. Other asymptotically valid tests based on efficient and inefficient estimators tend to over-reject when the null is correct, whereas rejection probabilities of the r_{CEL}-based test are close to the nominal level in their experiments. Kitamura, Tripathi, and Ahn (2004) note that the performance of their procedure is insensitive to the choice of bandwidth, and the standard cross-validation seems appropriate for selecting it automatically.

Smith (2003) extends the analysis of Kitamura, Tripathi, and Ahn (2004) by replacing the log-likelihood criterion with a Cressie-Read type divergence (see Section 3 for discussion on the Cressie-Read family of divergence). That is, he replaces (6.3) with

$$\sum_{j=1}^{n} w_{ij} \frac{2}{\alpha(\alpha+1)} \left[\left(\frac{p_{ij}}{w_{ij}} \right)^{-\alpha} - 1 \right]. \tag{6.6}$$

His estimator therefore includes the conditional empirical likelihood estimator as a special case where α is 0. This is analogous to the treatment of EL as a GMC presented in Section 3. Smith (2003) shows that his estimator for θ is first-order equivalent to $\hat{\theta}_{\text{CEL}}$. A related paper by Smith (2005) uses the GEL formulation to analyze the problem. It replaces $\log(1+y)$ in the dual form of log-likelihood contributions (6.4) with a general concave function. The resulting estimator corresponds to the dual formulation (3.16) of GMC, modified by kernel regression. It, therefore, further generalizes the Cressie-Read type estimator in Smith (2003) mentioned above. This line of research has been also pursued by Antoine, Bonnal, and Renault (2006). They study the conditional empirical likelihood procedure of Kitamura, Tripathi, and Ahn (2004), but replace the likelihood criterion in (6.3) with a chi-square distance. This corresponds to the criterion in (6.6) with $\alpha = -1$, and can be interpreted as a localized version of Euclidean likelihood discussed in Section 3. As noted there, this choice of criterion for GMC yields an explicit solution, which also applies to the localized version as well. This might lead to a modest saving of computational costs. Gagliardini, Gourieroux, and Renault (2004) develop a creative use of the Euclidean likelihood version of the conditional empirical likelihood procedure in option pricing.

An alternative empirical likelihood-based approach to estimate (6.1) has been suggested by Donald, Imbens, and Newey (2003). They use a series of functions of x to form a vector of instruments, and let the dimension of the vector increase as the sample size goes to infinity. Let $q^K(x)$ be such a vector, then the conditional moment restriction implies the unconditional moment restriction $E[g(z,\theta) \otimes q^K(x)] = 0$. Donald, Imbens, and Newey (2003) apply unconditional versions of empirical likelihood and GEL (hence GMC) to the implied unconditional moment condition model. Since the optimal GMM with instruments $\{a^*(x_i)\}_{i=1}^n$ is asymptotically efficient, if the vector q^K approximates a^* as $K \to \infty$, the optimal GMM applied to $g(z,\theta) \otimes q^K$ would be asymptotically efficient as well. A leading choice for q^K is spline: the s-th order spline with knots $t_1, ..., t_{K-3-1}$ is given by $q^K(x) = (1, x, ..., x^s, [(x - t_1 \vee 0)]^s, ..., [(x - t_{K-s-1}) \vee 0]^s)$. This means, however, that the dimension of the estimating function is high even for a moderate sample size. In view of the result by Newey and Smith (2004) described in Section 5.1, the two-step GMM for θ is likely to suffer from severe bias due to this high-dimensionality, whereas it is natural to expect empirical likelihood to perform well in a situation where the dimension of moments grows with the sample size. Donald, Imbens, and Newey (2003) develop asymptotic theory for the (generalized) empirical likelihood estimator for (6.1) under this environment. They show that this procedure also achieves the semiparametric efficiency bound $\mathcal{I}^{-1}$.

It is important to note that neither of the two EL-based procedures does not assume that the econometrician has *a priori* knowledge about a finite dimensional instrument vector a that identifies θ. Such knowledge is crucial for a two-step procedure discussed above. The treatment of the EL estimator in Donald, Imbens, and Newey (2003) imposes restrictions on the distribution of x, such as its density being bounded away from zero, and these are somewhat stronger that those assumed in Kitamura, Tripathi, and Ahn (2004). The two are quite different algorithmically. Maximization of the former is in some sense simpler than the latter, which requires calculation of local likelihood (6.4) at each x_i; on the other hand, the former involves construction of q^K, including choosing the basis functions and selection of the dimension of instruments K.

6.2. **Nonparametric specification testing.** A large literature exists for nonparametric goodness-of-fit testing for regression function. Let $y \in \mathbb{R}$ and $x \in \mathbb{R}^d$ be a pair of random elements, and consider the regression function $E[y|x] = m(x)$. Suppose one wishes to test a parametric specification $m(x) = m(x, \theta)$, such as a linear specification $m(x, \theta) = x'\theta$, against nonparametric alternatives. Many authors, including Eubank and Spiegelman (1990) and Härdle and Mammen (1993), consider this problem. See Hart (1997) for a review. A conventional approach to this problem is to compare predicted values from the parametric regression model and a nonparametric regression method in terms of an L_2-distance. The rest of this subsection discusses application of empirical likelihood to this problem. As the standard empirical likelihood ratio test possesses many desirable properties (see Sections 4 and 5), it is of interest to consider empirical likelihood-based nonparametric specification testing. It turns out that certain EL-based tests have asymptotic optimality properties.

The regression specification testing problem described above is a special case of the following. Consider an $\mathbb{R}^q-$valued parametric function $g(z, \theta), \theta \in \Theta \in \mathbb{R}^k$, of random variable z as in previous sections. Let x be a random variable such that under the null hypothesis

$$E[g(z, \theta)|x] = 0 \quad a.s. \text{ for some } \theta \in \Theta. \tag{6.7}$$

Letting $z = (y, x)$ and $g(z, \theta) = y - m(x, \theta)$, one obtains the regression specification hypothesis. The null hypothesis (6.7) is the identification restriction used in the previous section. Testing it, therefore, is equivalent to test the overidentifying restrictions in the model (6.1). This motivates the following test proposed by Tripathi and Kitamura (2003). It can be regarded as a nonparametric version of the likelihood ratio test.

Tripathi and Kitamura (2003) consider testing the conditional mean zero restriction over a compact set S, that is, $E[g(z, \theta)|x] = 0, \theta \in \Theta$ for $x \in S$. This is a common formulation used in the literature of nonparametric specification testing: see Aït-Sahalia, Bickel, and Stoker (2001), for example. With this in mind, define the conditional nonparametric log-likelihood function as a sum of localized nonparametric log-likelihood functions in (6.3) weighted by $t(x) = 1\{x \in S\}$:

$$\ell_{\mathrm{CNP}}(p_{11}, p_{12}, ..., p_{nn}) = \sum_{i=1}^{n} t(x_i)\ell_{\mathrm{LNP}i}(p_{i1}, ..., p_{in}).$$

Let ℓ^r_{CEL} signify the restricted maximum value of ℓ_{CNP} under the conditional moment constraints

$$\sum_{j=1}^{n} p_{ij} g(z_j, \theta) = 0, i = 1, ..., n, \theta \in \Theta. \tag{6.8}$$

Calculations similar to ones presented in the previous section show that[5].

$$\ell^r_{\mathrm{CEL}} = \sum_{i=1}^n t(x_i) w_{ij} \log w_{ij} + \sup_{\theta\in\Theta} \sum_{i=1}^n t(x_i) \min_{\gamma_i \in \mathbb{R}^q} - \sum_{j=1}^n w_{ij} \log(1 + \lambda_i' g(z_j, \hat{\theta}_{\mathrm{CEL}})).$$

It is straightforward to see that the maximum of ℓ_{CNP} without the restrictions (6.8) is attained at $p_{ij} = w_{ij}$, giving

$$\ell^u_{\mathrm{CEL}} = \sum_{i=1}^n t(x_i) \sum_{j=1}^n w_{ij} \log w_{ij}.$$

The log likelihood ratio statistic is therefore

$$\begin{aligned} r_{\mathrm{C}} &= -2(\ell^r_{\mathrm{CEL}} - \ell^u_{\mathrm{CEL}}) \\ &= 2 \inf_{\theta\in\Theta} \sum_{i=1}^n t(x_i) \max_{\gamma_i\in\mathbb{R}^q} \sum_{j=1}^n w_{ij} \log(1 + \lambda_i' g(z_j, \theta)). \end{aligned} \tag{6.9}$$

It is not essential that the second line in the definition of r_{C} is evaluated at θ that minimizes the expression. Other $\sqrt{n}$-consistent estimators for θ work without affecting the asymptotics.

Tripathi and Kitamura (2003) derive the limiting distribution of r_{C} under the null (6.7). Let S be $[0,1]^{\times d}$. This assumption is innocuous since it can be achieved by an appropriate transformation of x. Let $c_1(K) = \int K(u)^2 du$ be the roughness of the kernel function K used in (6.3). Define also $c_2(K) = \int [\int K(v)K(u-v)dv]^2 du$. Tripathi and Kitamura (2003) show the following: under the null hypothesis (6.7),

$$\frac{r_{\mathrm{C}} - h^{-d} q c_1(K)}{\sqrt{2h^{-d} q c_2(K)}} \xrightarrow{d} \mathrm{N}(0,1). \tag{6.10}$$

given that $d \leq 3$. Studentization of r_{C} for cases with $d > 3$ is possible, though more involved; see Tripathi and Kitamura (2003) for a formula that is valid for a general s. In practice, it seems that it is best to bootstrap r_{C} to obtain a reliable critical value for the empirical likelihood-based test (see simulation results in Tripathi and Kitamura (2003)), as critical values based on the first order approximation tend to lead to under-rejection. This is a phenomenon commonly observed throughout the literature of nonparametric specification testing; see, for example, Härdle and Mammen (1993).

The empirical likelihood test statistic r_{C} has features that distinguish it from other known procedures. To discuss these features of r_{C}, consider a large class of nonparametric testing procedures that includes standard tests such as Härdle and Mammen (1993) as a special case. For simplicity,

[5] Note that Tripathi and Kitamura (2003) uses a different normalization for p_{ij}'s, though this difference does not matter in implementing the test.

suppose both x and g are scalar-valued. Consider a weighted L_2-distance statistic based on kernel regression:

$$J_n(a) = h\sum_{i=1}^{n} a(x_i)\sum_{j=1}^{n}[w_{ij}g(z_j,\theta_n)]^2, a:[0,1]\to\mathbb{R}_+, \int a(x)^2dx = 1.$$

Note that the choice of the weighting function a is arbitrary. The statistic J_n can be standardized using the conditional variance function $\sigma^2(x) = \mathrm{var}(g(z_j,\theta_0)|x)$:

$$j_n(a) = \frac{h^{-1/2}[J_n(a) - c_1(K)\int \sigma a dx]}{\sqrt{2c_2(K)\int \sigma^2 a^2 dx}}.$$

The standardized weighted L_2-statistic $j_n(a)$ converges to a standard normal random variable in distribution under the null hypothesis. Note that the use of the kernel method above is inconsequential; other nonparametric methods yield results that are essentially the same.

The above asymptotic approximation result is valid under weak regularity conditions and can be used for testing the null hypothesis (6.7). The statistic $j_n(a)$, however, lacks an invariance property that r_{C} possesses. Let $b(x)$ be an arbitrary measurable function of x, and define $g^*(z,\theta) = b(x)g^*(z,\theta)$. Note that the null hypothesis (6.7) can be expressed in an alternative form $E[g^*(z,\theta)|x] = 0, \theta\in\Theta$ for every b. There is no mathematical or economic reason to prefer one parameterization over the other, since they are mathematically equivalent. Nevertheless, this reformulation affects the test since it essentially changes its weighting factor a. This dependence of empirical outcomes of the test on the formulation of the null hypothesis seems undesirable. In contrast, the empirical likelihood ratio statistic r_{C} is invariant with respect to the choice of b, since any change in b is absorbed into the variables $\lambda_i, i = 1,...,n$ in the definition of r_{C} in (6.9).

The empirical likelihood test based on r_{C} has an additional advantage in terms of its asymptotic power. Note that the tests based on r_{C} or $j_n(a)$ have nontrivial power against alternatives in an $n^{-1/2}h^{-1/4}$-neighborhood of the null when the dimension d of x is 1. To fix ideas, consider an alternative given by a function δ:

$$E[g(z,\theta_0)|x] = n^{-1/2}h^{-1/4}\delta(x), \qquad \theta\in\Theta.$$

Let f_x be the density function of x, and define

$$\mu(a,\delta) = \frac{\int_0^1 \delta^2 a f_x dx}{\sqrt{2c_2(K)\int_0^1 \sigma^2 a^2 dx}}.$$

The statistic $j_n(a)$ converges to $N(M(a,\delta),1)$ asymptotically under the sequence of alternatives. Let Φ denote the standard normal CDF, then if one chooses a critical value c for $j_n(a)$, the power function is given by

$$\pi(a,\delta) = 1 - \Phi(c - \mu(a,\delta)).$$

Tripathi and Kitamura (2003) show that the above results for $j_n(a)$ hold for the empirical likelihood ratio with a particular form of a:

$$a_{\mathrm{EL}}(x) = \frac{1}{\sigma(x)\int_0^1 \sigma^{-2}dx}.$$

Though the power function $\pi(a,\delta)$ can be maximized at $a^* = \text{const.} \times \delta^2(x) f_x(x)/\sigma^2(x)$ for a given δ, it depends on δ. Therefore such a procedure is infeasible in practice. One way to proceed is to integrate δ out from the power function $\pi(a,\delta)$ using a measure over the space of δ and to attempt maximizing the average power criterion. Wald (1943), facing a similar situation in a parametric multi-parameter testing problem, suggested the following. Suppose a hypothesis on a finite dimensional parameter takes the form $\theta = \theta_0 \in \mathbb{R}^p$. Let $\pi(\delta)$ denote the power function corresponding to a local alternative of the form $\theta_a = \theta_0 + n^{-1/2}\delta$. The average power function with weight P_δ is $\pi = \int \pi(\delta) dP_\delta$. Wald suggests using P_δ that is essentially the probability measure implied by the asymptotic distribution of the MLE for θ_0. Tripathi and Kitamura (2003) extend this principle to nonparametric specification testing. They suggest using the distribution of a continuous random function $\tilde{\delta}(x)$ on $s = [0,1]$ that mimics the distribution of the nonparametric conditional moment estimator

$$\hat{\delta}(x) = \frac{\sum_{i=1}^n K(\frac{x_i - x}{h}) g(x_i,\theta)}{\sum_{j=1}^n K(\frac{x_j - x}{h})}$$

to weight $\pi(a,\delta)$. Let $C([0,1])$ denote the set of continuous functions over $S = [0,1]$, then the random function $\tilde{\delta}(x)$ is a $C([0,1])-$valued random element. Let P_δ be the probability measure for the random function $\tilde{\delta}(x)$. The average power function is:

$$\pi(a) = \int_{C[0,1]} \pi(a,\tilde{\delta}) dP_\delta(\tilde{\delta}).$$

Calculus of variation shows that $\pi(a)$ is maximized at $a = a_{\mathrm{EL}}$ (Tripathi and Kitamura (2003)). That is, the empirical likelihood test is asymptotically optimal according to the average power criterion.

The methodology by Tripathi and Kitamura (2003) has been extended in various directions. Chen, Härdle, and Li (2003) investigate a method that is similar to the empirical likelihood test above. Their construction of empirical likelihood applies kernel regression to the function g, not likelihood. This also provides an asymptotically valid procedure. However, since it uses nonparametric regression of g on x, its finite sample behavior is not invariant with respect to multiplicative re-normalization of g with a function of x. This is in contrast to the methodology by Tripathi and Kitamura (2003), which has the invariance property since it smoothes the likelihood function rather than the moment function. On the other hand, Chen, Härdle, and Li (2003) develop asymptotic theory of their test for dependent processes. This extension is important, as they apply their method to test a defusion model for asset returns. Smith (2003) extends the test statistic r_{C} by replacing the log-likelihood criterion with a Cressie-Read type divergence measure. Smith's test, therefore, is also invariant against renormalization. It is natural to expect that his test possesses the average power optimality property of the r_{C}-based test.

There is a different empirical likelihood-approach for testing the conditional moment restriction (6.7). The insight that multiplying $g(z, \theta)$ by a growing number of appropriate instruments asymptotically imposes the conditional moment restrictions, used by Donald, Imbens, and Newey (2003) for obtaining an asymptotically efficient EL estimator, is valid here as well. Recall that the ELR function (2.9) evaluated at $\hat{\theta}_{\mathrm{EL}}$ serves as an appropriate test statistic for testing the overidentifying restrictions for the unconditional moment restriction model (2.2) with a fixed number of moment conditions. Testing (6.7) is asymptotically equivalent to testing a growing number of unconditional overidentifying restrictions of the form $E[g(z, \theta) \otimes q^K] = 0$ in the notation used in Section 6.1. It is, therefore, natural to use the maximum value of the ELR function calculated for the moment function $g \otimes q^K$ as a test statistic. Donald, Imbens, and Newey (2003) show that the test statistic

$$\frac{\sup_{\theta \in \Theta} \mathrm{elr}(\theta) - (qK - k)}{\sqrt{2(qK - k)}}$$

is distributed according to the standard normal distribution asymptotically under the null hypothesis (6.7), which corresponds to the limiting distribution for r_{C} obtained in (6.10).

6.3. **Dependent Data.**

6.3.1. *The Problem.* The foregoing sections explored empirical likelihood methods with independent observations. This section points out various impacts of dependence on the results discussed so far in the current paper, and presents empirical likelihood-based approaches that are suitable for time series data. The introduction of dependence typically necessitates modification of the definition of empirical likelihood. This should be clear from the form of the nonparametric log likelihood (2.1), which is interpreted as the log-likelihood of independent multinomial data. Some of the standard asymptotic properties of empirical likelihood no longer hold under dependence without appropriate modifications. To see this point, suppose stationary and weakly dependent time series observations $\{z_t\}_{t=1}^T$ are given. (See Kitamura (1997) for some discussions on the notion of weak dependence.)

Consider an unconditional moment condition model as in Section 2:

$$E[g(z_t, \theta_0)] = 0, \quad \theta_0 \in \Theta. \tag{6.11}$$

The only difference from (2.2) is that z_t is a dependent process in (6.11). Recall that the standard EL estimator solves the FOC of the form (5.3). As $\hat{D}(\hat{\theta}_{\mathrm{EL}})$ and $\hat{S}(\hat{\theta}_{\mathrm{EL}})$ are weighted averages, they converge to their population counterparts $D = E[\nabla_\theta g(z_t, \theta_0)]$ and $S = E[g(z_t, \theta_0)g(z_t, \theta_0)']$ in probability. Expanding (5.3) with respect to $\hat{\theta}_{\mathrm{EL}}$ in $\bar{g}$ around θ_0 and solving for $\hat{\theta}_{\mathrm{EL}} - \theta_0$,

$$\sqrt{T}(\hat{\theta}_{\mathrm{EL}} - \theta_0) = (D'S^{-1}D)^{-1}DS^{-1}\sqrt{T}\bar{g}(\theta_0) + o_p(1), \quad \bar{g}(\theta_0) = T^{-1}\sum_{t=1}^{T} g(z_t, \theta_0). \tag{6.12}$$

Under a mild mixing condition, such as the one used in Kitamura (1997), the term $\sqrt{T}\bar{g}(\theta_0)$ follows the central limit theorem:

$$\sqrt{T}\bar{g}(\theta_0) \xrightarrow{d} \mathrm{N}(0, \Omega), \quad \Omega = \sum_{j=-\infty}^{\infty} E[g(z_t, \theta_0)g(z_{t+j}, \theta_0)'],$$

yielding

$$\sqrt{T}(\hat{\theta}_{\mathrm{EL}} - \theta_0) \xrightarrow{d} \mathrm{N}(0, (D'S^{-1}D)^{-1}D'\Omega D(D'S^{-1}D)^{-1}).$$

What this means is that the EL estimator $\hat{\theta}_{\mathrm{EL}}$ is asymptotically first-order equivalent to the GMM estimator with a sub-optimal weighting matrix ($=S^{-1}$), whereas Ω^{-1} should be used for optimal weighting. The standard empirical likelihood therefore yields an estimator that is $T^{1/2}-$ consistent and obeys a normal law asymptotically, but it is less efficient than the optimally weighted GMM estimator. This fact also affects EL-based inference. Suppose one is interested in testing the hypothesis $\theta_0 \in \Theta_0 \subset \Theta$. Arguments as above show that the ELR statistic defined in (2.8) is asymptotically equivalent to the difference between the minimum values of the quadratic form $T\bar{g}(\theta)S^{-1}\bar{g}(\theta)$ with and without the constraint $\theta \in \Theta_0$. Since the quadratic form is not weighted by the appropriate matrix Ω^{-1}, the ELR statistic fails to converge to the desired chi-square random variable with the degrees of freedom $\dim(\Theta) - \dim(\Theta_0)$.

There are some possible approaches to solve the above problems. The first uses a parametric model, and the second avoids such parametric modeling. The third strikes a middle ground between the two approaches. The fourth uses a spectral method, which has a limited range of applications relative to the other methods.

6.3.2. *Parametric Approach.* A rather obvious method to deal with dependence is to introduce a parametric model to remove dependence in the data. For example, one may use a $p-$th order (vector) autoregressive model for the purpose. Let L denote the backshift operator, i.e. $Lx_t = x_{t-1}$. Consider a $p-$th order polynomial $B(L;\xi)$ parameterized by a finite dimensional vector $\xi \in \Xi \subset \mathbb{R}^J$. Suppose operating $B(L,\xi_0)$ to $g(x_t,\theta_0)$ yields $\epsilon_t = B(L,\xi_0)g(x_t,\theta_0)$ which is a martingale difference sequence (mds). Define $z_t^* = (z_t, z_{t-1}, ..., z_{t-p})$, $t = p+1, ..., T$ and $\theta^* = (\theta', \xi')' \in \Theta^* = \Theta \times \Xi$. Then the function $g^*(z_t^*, \theta^*) = [B(L,\xi)g(z_t,\theta)] \otimes [1, g(z_{t-1},\theta)', ..., g(z_{t-p},\theta)']'$ satisfies moment restrictions, i.e. $E[g^*(z_t^*,\theta^*)] = 0$ at $\theta^* = \theta_0^* = (\theta_0', \xi_0')'$. If such θ^* is unique, application of EL to g^* is justified. Moreover, the sequence $\{g^*(z_t^*,\theta_0^*)\}_{t=1}^T$ is also a mds by construction. An application of the martingale difference CLT to $\bar{g}^*(\theta_0^*) = T^{-1}\sum_{t=p+1}^T g(z_t^*, \theta_0^*)$ yields

$$\sqrt{T}\bar{g}^*(\theta_0,\xi_0) \xrightarrow{d} \mathrm{N}(0,S^*), \quad S^* = E[g^*(z_t^*,\theta_0^*)g^*(z_t^*,\theta_0^*)'].$$

The standard empirical likelihood estimator in Section 2 applied to g^* yields an appropriate estimator $\hat{\theta}_{\mathrm{EL}}^*$, in the sense that

$$\sqrt{T}(\hat{\theta}_{\mathrm{EL}}^* - \theta_0^*) \xrightarrow{d} \mathrm{N}(0,(D^{*\prime}S^{*-1}D^*)^{-1}), \quad D^* = E[\nabla_{\theta^*}g^*(z_t^*,\theta_0^*)]. \tag{6.13}$$

This gives the joint limiting distribution for the empirical likelihood estimator for the parameter of interest θ_0 and the nuisance parameter ξ_0. Suppose the $B(L,\xi)$ is parameterized as an unrestricted $p-$th order VAR, i.e.

$$B(L,\xi) = I_q - \sum_{j=1}^p \Xi_j L^j, \xi = (vec\Xi_1', ..., vec\Xi_p')'$$

Write $\hat{\theta}_{\mathrm{EL}}^* = (\hat{\theta}_{\mathrm{EL}}', \hat{\xi}_{\mathrm{EL}}')'$, then a calculation shows that the marginal limiting distribution for $\hat{\theta}_{\mathrm{EL}}$ in the joint distribution given in (6.13) is:

$$\sqrt{T}(\hat{\theta}_{\mathrm{EL}}^* - \theta_0) \xrightarrow{d} \mathrm{N}(0,(D'\Omega^{-1}D)^{-1}).$$

Therefore $\hat{\theta}_{\mathrm{EL}}^*$ achieves the same asymptotic efficiency as the optimally weighted GMM for (6.11). It appears that the method described above to estimate the model (6.11) is new, though some researchers considered procedures related to the above methodology. They focus on time series models such as an AR model rather than a structural model (6.11), therefore $g(z_t,\theta) = z_t$. They are concerned with inference for the parameter ξ_0 in $B(L,\xi_0)$. One can use a sieve method where the order p of the polynomial $B(L,\xi)$ goes to infinity as the sample size T grows (e.g. Bravo (2005b)).

6.3.3. *Nonparametric Approach.* The above approach has some aspects that are not entirely satisfactory, as it relies on the parametric filter $B(L, \xi)$. This reduces the appeal of empirical likelihood as a nonparametric procedure. It also involves joint estimation of the parameter of interest θ_0 and the nuisance parameters ξ_0, which can be high-dimensional for a moderate or large p. Fortunately, it is possible to treat dependence fully nonparametrically in empirical likelihood analysis without relying on a time series model as used above. The idea is to use blocks of consecutive observations to retrieve information about dependence in data nonparametrically, therefore termed blockwise empirical likelihood (BEL). This is the approach first proposed by Kitamura (1997) and Kitamura and Stutzer (1997). There is interesting parallelism between the bootstrap and empirical likelihood as pointed out, among others, by Hall and LaScala (1990), and this is no exception: BEL is closely related to the blockwise bootstrap proposed by Hall (1985), Carlstein (1986) and Künsch (1989).

Implementation of blockwise empirical likelihood proceeds as follows. Again, consider (6.11) where $\{z_t\}_{t=1}^T$ are weakly dependent. The first step is to form blocks of observations. The following description specializes to the "fully overlapped blocking" scheme in the terminology of Kitamura (1997), for the sake of simplicity. This is essentially equivalent to the "time-smoothing" method proposed by Kitamura and Stutzer (1997). A general blocking scheme that includes overlapping blocking and non-overlapping blocking as two extreme special cases is discussed in Kitamura (1997). The first step is to form data blocks: the $t-$th block of observations is given by $B_t = (z_t, z_{t+1}, ..., z_{t+M-1})$, $t = 1, ..., T - M + 1$ for an integer M. Suppose $M \to \infty$ and $M = o(T^{1/2})$ as $T \to \infty$. The purpose of blocking is to retain the dependence pattern of z_t's in each block B_t of length M. Since M grows slowly as the sample size grows, it captures information about weak dependence in the data asymptotically in a fully nonparametric way. The second step is to calculate what Kitamura and Stutzer (1997) call the $t-$th "smoothed moment function":

$$\psi(B_t, \theta) = M^{-1} \sum_{s=1}^{M-1} g(z_{t+s}, \theta).$$

The third step is to apply the empirical likelihood procedure discussed in Section 2 with $g(z_i, \theta), i = 1, ..., n$ with $\psi(B_t, \theta), t = 1, ..., T-M+1$. Proceeding as before yields the blockwise empirical likelihood function profiled at θ:

$$\ell_{\text{block}}(\theta) = \min_{\gamma \in \mathbb{R}^q} - \sum_{t=1}^{T-M+1} \log(1 + \gamma' \psi(B_t, \theta))) - (T - M + 1) \log(T - M + 1). \tag{6.14}$$

This corresponds to (2.5). Let $\hat{\theta}_{\text{block}}$ denote the maximizer of ℓ_{block} over Θ; this is the blockwise empirical likelihood estimator of Kitamura (1997). Kitamura (1997) shows that the BEL estimator has the following limiting distribution

$$\sqrt{T}(\hat{\theta}_{\text{block}} - \theta_0) \xrightarrow{d} \mathrm{N}(0, (D'\Omega^{-1}D)^{-1}). \tag{6.15}$$

The blockwise empirical likelihood incorporates information about dependence in the estimator. It achieves the same asymptotic efficiency as an optimally weighted GMM in a fully nonparametric way, but it avoids preliminary estimation of the optimal weighting matrix Ω^{-1}. The latter fact means that the BEL estimator shares some advantages with the standard EL estimator, such as its invariance property with respect to re-normalization of the moment condition vector g.

It is easy to see how the "right" asymptotic distribution in (6.15) is achieved by BEL. BEL replaces the original moment function g with ψ. An approximation similar to (6.12) holds after this replacement, but the relationships $E[\nabla_\theta \psi(z_t, \theta_0)] = D$ and $\lim_{M\to\infty} E[M\psi(B_t, \theta_0)\psi(B_t, \theta_0)'] = \Omega$ imply that

$$\sqrt{T}(\hat{\theta}_{\text{block}} - \theta_0) = (D'\Omega^{-1}D)^{-1}D\Omega^{-1}\sqrt{T}\bar{\psi}(\theta_0) + o_p(1), \tag{6.16}$$
$$\bar{\psi}(\theta_0) = (T - M + 1)^{-1} \sum_{t=1}^{T-M+1} \psi(B_t, \theta_0).$$

Noting that $T^{1/2}\bar{\psi}(\theta_0) \xrightarrow{d} \mathrm{N}(0, \Omega)$, (6.15) follows. This argument shows that BEL implicitly "estimates" $\Omega = \lim_{M\to\infty} E[M\psi(B_t, \theta_0)\psi(B_t, \theta_0)']$ by its sample counterpart. The "correct" weighting matrix $\Omega = \sum_{-\infty}^{\infty} E[g(z_t, \theta_0)g(z_t, \theta_0)']$ emerges in the approximation (6.16) from the probability limit of a weighted average of $M\psi(B_t, \theta_0)\psi(B_t, \theta_0)'$, $t = 1, ..., T - M + 1$. Note that the normalized sample variance of the form

$$\hat{\Omega} = (T - M + 1)^{-1} \sum_{t=1}^{T-M+1} M\psi(B_t, \theta_0)\psi(B_t, \theta_0)'.$$

corresponds to the Bartlett kernel estimator of the "long-run covariance matrix" as proposed by Newey and West (1987) with lag length $M - 1$. This suggests that one may select M in BEL by using a lag-selection rule developed in the literature of long-run covariance estimation (see, for example, Andrews (1991) and Priestley (1981)). The above observation also implies that it is possible to use weights other than the flat weighting by M^{-1} in calculating ψ above. Such alternative weights correspond to other kernel estimators of long-run covariances, as pointed out by Kitamura and Stutzer (1997) and Kitamura (1997). See Smith (2004) for a comprehensive account of this correspondence in the context of GEL.

The blockwise empirical likelihood function ℓ_{block} also yields a likelihood ratio test statistic to which standard asymptotic results apply. Suppose one is interested in testing a hypothesis $\theta_0 \in \Theta_0$

with $\dim(\Theta_0) = k - s$. Let

$$r_{\text{block}} = -2c_T^{-1}(\sup_{\theta\in\Theta_0} \ell_{\text{block}}(\theta) - \sup_{\theta\in\Theta} \ell_{\text{block}}(\theta)), \quad c_T = \frac{(T-M+1)M}{T}.$$

The factor c_T is necessary to account for the fact that blocks are overlapping, in the following sense. There are $T - M + 1$ blocks of observations and each data block B_t consists of M observations $z_t, ..., z_{t-M+1}$, therefore seemingly $(T - M + 1)M$ observations enter the likelihood function ℓ_{block}. But the actual number of observation is T, thus c_T measures how many times each observation gets double-counted. The above likelihood ratio statistic with the correction term c_T^{-1} converges to a chi-squared random variable with s degrees of freedom in distribution under the null hypothesis. Similarly, modify the definition of the ELR function in (2.9) to define

$$\text{elr}_{\text{block}}(\theta) = -2c_T^{-1}[\ell_{\text{block}}(\theta) + (T - M + 1)\log(T - M + 1)].$$

It can be shown that $\text{elr}_{\text{block}}(\theta_0) \xrightarrow{d} \chi^2_s$, which can be used for a test that is analogous to the Anderson-Rubin test. Likewise, the value of $\text{elr}_{\text{block}}$ at the BEL estimator asymptotically obeys the χ^2_{q-k} law; therefore, it offers a test for the overidentifying restrictions of the moment condition model (6.11) for time series data. Kitamura (1997) also extends the Bartlett correctability result by DiCiccio, Hall, and Romano (1991) of $\text{elr}(\theta_0)$ for iid data to $\text{elr}_{\text{block}}(\theta_0)$ with weakly dependent data. Let a denote the Bartlett factor (see Kitamura (1997) for its expression). The result obtained in the paper shows that adjusting $\text{elr}_{\text{block}}(\theta_0)$ by a improves the accuracy of the chi-square approximation for the distribution of the test statistic from

$$\Pr\{\text{elr}_{\text{block}}(\theta_0) \le x\} = \Pr\{\chi^2_q \le x\} + O(T^{-2/3})$$

to

$$\Pr\{\text{elr}_{\text{block}}(\theta_0)(1 + T^{-1}a) \le x\} = \Pr\{\chi^2_q \le x\} + O(T^{-5/6}).$$

Kitamura (1997) also shows that the idea of using blocks of data to construct empirical likelihood can be extended to inference for infinite dimensional parameters, such as the spectral density of z_t.

The blockwise empirical likelihood method has been extended in various directions. Nordman, Sibbertsen, and Lahiri (2006) make an interesting discovery by considering inference for the mean $E[z_t]$ when the process z_t exhibits the so-called long range dependence behavior. It is known in the literature that blocking methods used in the bootstrap and subsampling tend to break down under long range dependence. A long range dependent process can be characterized by how slow its autocovariance function decays, or, alternatively by the behavior of its spectral density at the origin. Using the latter formulation, suppose the spectral density $f_z(\omega)$ of the process $\{z_t\}_{t=-\infty}^{\infty}$ at frequency ω is of the same order as $|\omega|^{-2d}$, $d \in (-\frac{1}{2}, \frac{1}{2})$. A non-zero value of d corresponds to long-range dependence. The essence of the discovery of the Nordman, Sibbertsen, and Lahiri (2006) is that the blockwise empirical likelihood function as defined by Kitamura (1997) can be modified suitably for long range dependent cases. In particular, they show that the adjustment factor $c_T = \frac{(T-M+1)M}{T}$ used for the weakly dependent case needs to be modified as follows:

$$c_{T,n} = (T - M + 1)\left(\frac{M}{T}\right)^{1-2d}.$$

The value of d is zero if z_t is weakly dependent. In this case $c_{0,T} = c_T$ and the factor reduces to the one proposed by Kitamura (1997).

Kitamura and Stutzer (1997) apply a variant of the blocking scheme as above to develop an exponential tilting estimator for weakly dependent processes. Indeed, a number of subsequent papers that study various aspects of BEL have appeared. Smith (1997) notes that Kitamura and Stutzer's blocking method remains valid for the entire GEL family; hence the same is expected to hold for the GMC family in Section 3 as well. Higher order properties of BEL and other blockwise versions of GEL estimators, in the spirit of Newey and Smith (2004), are investigated by Anatolyev (2005). Bravo (2005a) studies the application of the blocking-based method by Kitamura (1997) as well as the blocking-after-prewhitening method by Kitamura (1996a) described in the next section to the saddle-point exponential tilting estimator by Kitamura and Stutzer (1997). Likewise, Lin and Zhang (2001) and Bravo (2002) replace empirical likelihood in (6.14) with Euclidean likelihood and the Cressie-Read divergence, respectively, and confirm that the first order asymptotic results for estimation and testing derived for BEL in Kitamura (1997) still hold.

You, Chen, and Zhou (2006) find a different application of BEL to what essentially is a random effects model for longitudinal data. In their application a block is formed per individual, so the length of each block is equal to the number of observations available for each individual. Therefore it does not go to infinity in their asymptotics. This method is robust against heteroskedasticity and within-group correlation. You, Chen, and Zhou (2006) report experimental results that indicate that BEL tends to produce much shorter confidence intervals than others with comparable coverage probabilities, such as those obtained based on normal approximations with robust standard errors. This fact is consistent with the optimal power properties of the empirical likelihood ratio test as outlined in Sections 4.2 and 4.3. Zhang (2006) applies BEL to NA (negatively associated) time series (see Joag-Dev and Proschan (1983) for a definition of an NA process) and proves its asymptotic validity. Allen, Gregory, and Shimotsu (2005) propose a bootstrapping procedure based on BEL. The idea is to extend the

EL-based bootstrap method by Brown and Newey (2002), which will be described in Section 8, to dependence data using BEL.

6.3.4. *A Middle Ground.* The blocking approach avoids arbitrary specification of dynamics. Therefore, the empirical likelihood function is obtained without imposing restrictions other than the model restriction (6.11). When observations are highly persistent, however, in the sense that the autocorrelation function decays slowly as the number of lags increases, it might take long blocks to capture dependence in the data. This requires the data size to be large. If the data process under consideration appears highly persistent and the size of the available data set is small, the blocking approach might need a modification. One possibility is to merge the parametric and blocking approaches, borrowing the idea of prewhitening from the literature of spectral density estimation. See Section 7.4.1 of Priestley (1981) for the pre-whitening method and Andrews and Monahan (1992) for an application in econometrics. The idea of prewhitening in the spectral analysis of highly persistent processes is as follows. First, fit a lower order (vector) autoregressive model to the original series and use it as a filter to reduce its dependence, so that the process is closer to white noise after filtering. This is the prewhitening step. Second, apply a standard spectral method to estimate the spectrum of the prewhitened process. Third, use the coefficients of the (V)AR model used in prewhitening to obtain an estimate of the spectrum of the original process. This last step is called re-coloring.

Applications of prewhitening in empirical likelihood have been investigated by Kitamura (1996a). Consider the model (6.11) once again, where z_t is highly persistent. As before, apply a parametric VAR model $B(L,\xi)$ to filter the process $\{g(z_t,\theta)\}_{t=1}^T$, though $B(L,\xi)$ is not meant to be the true model that generates the process $g(z_t,\theta_0)$. Therefore, the filtered process $B(L,\xi)g(z_t,\theta)$ would exhibit a certain degree of dependence for every value of $\theta^* = (\theta', \xi')'$; in particular, it is not supposed to be a mds. The purpose of the filter $B(L,\xi)$ is to reduce the dependence in the process, not eliminating it. A low order filter, even a first order model, may suffice for the purpose. Such a choice avoids the problem of overparameterization, which can be a serious issue in the purely parametric approach of Section 6.3.2. Now let $z_t^* = (z_t, ..., z_{t-p})$ and apply the blocking technique described above to the process $g^*(z_t^*, \theta, \xi) = [B(L,\xi)g(z_t,\theta)] \otimes (1, g(z_{t-1},\theta)', ..., g(z_{t-p},\theta)')'$ to deal with the dependence not captured by the filter:

$$\psi^*(B_t^*, \theta, \xi) = M^{-1}\sum_{s=1}^{M-1} g^*(z_{t+s}^*, \theta, \xi), \quad B_t^* = (z_t^*, ..., z_{t+M-1}^*), \quad t = p+1, ..., T-M+1.$$

The blockwise empirical log-likelihood with prewhitening is

$$(6.17)\ \ \ell_{\text{pwblock}}(\theta) = \sup_{\xi\in\Xi}\min_{\gamma\in\mathbb{R}^{q+\dim\Xi}} - \sum_{t=p+1}^{T-M+1}\log(1+\gamma'\psi^*(B_t^*,\theta,\xi))-(T-M-p+1)\log(T-M-p+1).$$

Let $\hat{\theta}_{\text{pwblock}}$ denote the maximizer of $\ell_{\text{pwblock}}(\theta)$. The block length parameter M needs to go to infinity such that $M = o(T^{1/2})$, as assumed for BEL. Since the filter $B(L,\xi)$ is not a model, $\hat{\xi}_{\text{pwblock}}$ would converge to some "pseudo-true" value and therefore its asymptotic behavior is not of main interest. Regarding the parameter of interest θ_0, the following holds

$$\sqrt{T}(\hat{\theta}_{\text{pwblock}}-\theta_0)\xrightarrow{d}\mathrm{N}(0,(D'\Omega^{-1}D)^{-1}).$$

To carry out inference, replace $\ell_{\text{block}}(\theta)$ with $\ell_{\text{pwblock}}(\theta)$ in the definitions of r_{block} and $\text{elr}_{\text{block}}(\theta), \theta\in\Theta$. The resulting prewhitened versions r_{pwblock} and $\text{elr}_{\text{pwblock}}(\theta), \theta\in\Theta$ have the same asymptotic properties as their BEL counterparts without prewhitening. An interesting feature of this procedure is that there is no need to apply recoloring explicitly, since it is done implicitly in the empirical likelihood algorithm.

6.3.5. *Frequency Domain Approach.* Yet another approach to deal with dependence in the empirical likelihood analysis is to apply frequency domain methods, as proposed by Monti (1997). This work follows the Whittle likelihood methodology, therefore a parametric model for the spectral density (e.g. the spectral density function implied by a parametric ARMA model) is considered. The method is suitable for parametric time series models and differs from the block-based methodologies discussed in Sections 6.3.3 and 6.3.4, where the goal is to treat dependence nonparametrically. Nordman and Lahiri (2004) shows that the frequency domain empirical likelihood applies to a class of statistics termed ratio statistics (see Dahlhaus and Janas (1996)), allowing possible long range dependence.

6.4. **Further Applications of EL.** An interesting aspect of empirical likelihood is that it allows the researcher to combine information from two data sets in a natural manner. Chen, Leung, and Qin (2003) discuss an application of empirical likelihood when a complete data set (validation set) and a data set that includes covariates and surrogates (non-validation set) are available. They find that their empirical likelihood based method yields highly accurate confidence intervals. Hellerstein and Imbens (1999) discuss the use of empirical likelihood to estimate a regression model when information on some moments is available from auxiliary data.

Wang and Rao (2002) develop a valid empirical likelihood ratio test in a missing data problem, where nonparametric imputation is used under the missing at random (MAR) assumption. They also derive an empirical likelihood estimator that incorporates information in additional moment conditions. Tripathi (2005) considers EL-based estimation of models with stratified data.

Empirical likelihood has been applied to the problem of weak instruments; see Caner (2003), Guggenberger and Smith (2005) and Otsu (2006). These papers use empirical likelihood or GEL mainly in an LM-test setting (which, if appropriately defined, is known to work for the weak IV problem; see Kleibergen (2002)). This is a rather tricky problem for EL, because many distinctive properties of empirical likelihood-based inference crucially depend on the structure of the empirical likelihood ratio test statistic, and they do not generally carry over to LM-type tests.

7. Misspecification

It is well-known that OLS and parametric MLE yield results that can be regarded as best approximations, where the criteria are mean square error (White (1980)) and the Kullback-Leibler divergence (see Akaike (1973) and White (1982)), respectively. Such a best approximation result does not carry over to the two-step GMM, since the probability limit of the two-step GMM depends on the weighting matrices used in both steps in a complicated and uninterpretable way. A one-step estimation with a sub-optimal weighting matrix may avoid this issue, but such an estimator loses efficiency when the model is correctly specified.

Interestingly, the GMC estimator possesses an approximation property analogous to that of MLE. To illustrate this point, it is useful to summarize basic results from the theory of parametric ML. Suppose $\{z_i\}_{i=1}^n \sim_{iid} \mu$. The econometrician uses a finite dimensional vector $\xi \in \Xi$ for parameterization, so the model is given by $\mathcal{P}_{\text{par}} = \{P_\xi | \xi \in \Xi\}$. The model $\mathcal{P}_{\text{par}}$ is misspecified if it does not contain μ. MLE then converges to ξ^* such that $P_{\xi^*} = \text{argmin}_{P \in \mathcal{P}_{\text{par}}} K(P, \mu)$. That is, MLE finds the probability measure that is closest to the true distribution μ in terms of the KL divergence.

Now, consider the moment condition model (2.2). This means that the statistical model is given by $\mathcal{P}$ defined in (3.2), instead of $\mathcal{P}_{\text{par}}$ above. Suppose $\mathcal{P}$ is misspecified. A useful fact is that a GMC estimator finds the best approximation for the true measure μ, where the approximation criterion is given by the contrast function (3.1). For example, Kitamura (1998) studies the asymptotic behavior of the exponential tilt saddle-point estimator (3.14). The main results are as follows. $\hat{\theta}$ that solves (3.14) and $\widehat{\bar{P}(\hat{\theta})}(A)$ in (3.15) converge to values θ^* and $P^* = \bar{P}(\theta^*)$, where

$$K(P^*, \mu) = \inf_{p \in \mathcal{P}} K(P, \mu).$$

That is, the values θ^* and P^* correspond to the solution of the approximation problem $\inf_{P\in\mathcal{P}} K(P,\mu)$. The exponential tilt estimator and the corresponding probability measure estimator are sample analog of $(\theta^*, \bar{P}(\theta^*))$. Imbens (1997) makes a related observation that the empirical likelihood estimator minimizes the Kullback-Leibler divergence between the empirical distribution and the probability distribution under the moment constraint; see also Chen, Hong, and Shum (2001). These results are expected to extend to other members of GMC.

Recall that some GEL estimators can be interpreted as dual GMC estimators when ρ in (3.17) is the convex conjugate of a Cressie-Read divergence. It is therefore obvious that the best approximation result holds for the corresponding subset of the GEL family. In general, however, GEL estimators may not be represented as GMC. For this reason some GEL's do not seem to provide best approximation interpretations presented here.

Kitamura (1998) also derives the asymptotic distribution of the exponential tilting estimator under misspecification and extends Vuong's model comparison test (Vuong (1989)). Vuong's original test is concerned with likelihood-ratio testing between two non-nested parametric models. His model comparison measure for two parametric models $\mathcal{P}_{\text{par}}$ and $\mathcal{Q}_{\text{par}}$ is

$$\delta = \inf_{P\in\mathcal{P}_{\text{par}}} K(\mu,P) - \inf_{Q\in\mathcal{Q}_{\text{par}}} K(\mu,Q). \tag{7.1}$$

Vuong (1989) shows that a normalized likelihood ratio statistic ($= LR$) converges to δ in probability. This can be used to test the null hypothesis $\delta = 0$, since $\sqrt{n}LR/s_\delta$, where s_δ is an appropriate studentization factor, converges to the standard normal distribution.

Kitamura (1998) shows that this idea works for moment condition models. The motivation of the paper comes from the notion that many economic models are best viewed as approximations. Even though moment condition models are considered to be more robust than parametric models, they still often come from highly stylized economic models. The researcher needs to confront the issue of misspecification in such a situation. For example, a leading example of applications of moment condition models is the classic study of asset pricing models by Hansen and Singleton (1982). If one subjects such a model to specification tests, oftentimes negative results emerge, implying potential misspecification of the model. Moreover, there are many other non-nested moment conditions implied by different asset pricing models, such as cash-in-advance models. It is then of interest to compare two potentially misspecified competing moment condition models.

Consider two moment condition conditions that are non-nested, and possibly misspecified:

$$E[g_1(z,\theta_1)] = 0, \theta_1 \in \Theta_1 \quad \text{and} \quad E[g_2(z,\theta_2)] = 0, \theta_2 \in \Theta_2.$$

From these conditions define two sets of probability measures as in (3.2). Call them $\mathcal{P}_1$ and $\mathcal{P}_2$. Suppose the researcher decides to use a contrast function of the form (3.1) to measure the divergence between the true probability measure and the probability measures implied by the model. Discussions in Section 3.1 imply that different choices of ϕ generate a wide variety of criteria. Once a criterion is chosen, define

$$\begin{aligned}\delta &= \inf_{P_1\in\mathcal{P}_1} D(P_1\|\mu) - \inf_{P_2\in\mathcal{P}_2} D(P_2\|\mu)\\ &= \inf_{\theta_1\in\Theta_1} v_1(\theta_1) - \inf_{\theta_2\in\Theta_2} v_2(\theta_2)\end{aligned}$$

where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ are the unknown parameters in the two moment condition models, and v_1 and v_2 are the corresponding value functions (see Equation (**P**) in Section 3). $\delta = 0$ means that Models $\mathcal{P}_1$ and $\mathcal{P}_2$ are equally good approximations of the true probability measure μ in terms of the researcher's criterion function. Likewise, a positive (negative) δ implies that $\mathcal{P}_2$ ($\mathcal{P}_1$) fits to the true data distribution better than P_1 (P_2) does. Using the sample dual form (3.9), define:

$$\begin{aligned}\hat{\delta} &= \inf_{\theta_1\in\Theta_1}\max_{\lambda_1\in\mathbb{R},\gamma_1\in\mathbb{R}^q}\left[\lambda_1 - \frac{1}{n}\sum_{i=1}^n \phi^\star(\lambda_1+\gamma_1' g_1(z_i,\theta_1))\right]\\ &\quad - \inf_{\theta_2\in\Theta_2}\max_{\lambda_2\in\mathbb{R},\gamma_2\in\mathbb{R}^q}\left[\lambda_2 - \frac{1}{n}\sum_{i=1}^n \phi^\star(\lambda_2+\gamma_2' g_2(z_i,\theta_2))\right].\end{aligned} \tag{7.2}$$

Kitamura (1998) considers the asymptotic distribution of $\hat{\delta}$ for the case where the contrast function is given by the Kullback-Leibler divergence $K(\cdot,\mu)$, therefore the resulting estimators are the exponential tilt estimators (3.14). As in Vuong's test, with an appropriate scaling factor $s_\delta > 0$, the asymptotic distribution of the statistic under the null hypothesis $\delta = 0$ is standard normal:

$$\frac{n^{1/2}\hat{\delta}}{s_\delta} \xrightarrow{d} \mathrm{N}(0,1).$$

Kitamura (1998) also discusses how to estimate s_δ. See Christoffersen, Hahn, and Inoue (2001) for an application of this model comparison test to value-at-risk modeling.

Chen, Hong, and Shum (2001) note that such an idea can be also used to carry out model comparison between a parametric specification and a semiparametric one. In their case, $\mathcal{P}_1$ comes from a parametric model: the researcher parameterizes the distribution of data z as $P_\xi, \xi \in \Xi$, so $\mathcal{P}_1 = \{P_\xi | \xi \in \Xi\}$. This model is compared against a moment condition model, which generates $\mathcal{P}_2$. Chen, Hong, and Shum (2001) use $D(P,\mu) = \int \log \frac{d\mu}{dP} d\mu$ as the contrast function.

Kitamura (2002) also explores the issue of model comparison tests, focusing on the case where there exist covariates. The paper considers estimation of conditional moment restriction models of the form $E[g(z,\theta)|x] = 0$. For example, suppose one considers modeling the conditional mean of a variable y given x using a parametric function, e.g. $E[y|x] = \beta' x$. Alternatively, a median regression $\mathrm{med}(y|x) = \beta' x$ can be used. These models are non-nested, therefore conventional testing methods do not work for testing one against the other. Note also that the presence of covariates is crucial in these models. A typical choice of instrumental variables for these models is x, but moment conditions such as $E[x(y - \beta' x)] = 0$ impose a just-identifying restriction so that the corresponding set of measures $\mathcal{P}$ always include μ. GMC-based testing methods discussed above therefore do not work, since $\inf_{P \in \mathcal{P}} D(P, \mu) = 0$ no matter how badly the regression function is specified. It is of course possible to add more instruments to obtain over-identifying restrictions, but such a method involves an *ad hoc* choice of instruments. A solution to this problem by Kitamura (2002) is to impose the conditional moment restriction directly and apply GMC by following the methodology in Kitamura, Tripathi, and Ahn (2001) described in Section 6. Kitamura (2002) also develops asymptotic theory for misspecified quantile regression models. This is a topic that has attracted attention in the recent literature (see Kim and White (2002) and Angrist, Chernozhukov, and Fernandez (2005)). For example, Angrist, Chernozhukov, and Fernandez (2005) investigate the asymptotic behavior of the linear quantile regression estimator of Koenker and Bassett (1978). In contrast, Kitamura (2002) considers the asymptotics of the estimator in Kitamura, Tripathi, and Ahn (2001) and provides a best approximation characterization. The method in Kitamura (2002) is also useful in evaluating and comparing a parametric model with covariates with mean/quantile regression models.

The consideration on misspecification also raises an interesting issue about robustness. If one accepts the view that the econometric model under consideration (e.g. the moment condition model (2.2)) is a reasonable yet misspecified approximation of the unknown true structure, it may be desirable to use an estimator that is robust to misspecification. Roughly speaking, there are two issues involved in assessing the robustness of an estimator. One is about the bias of the estimator due to the misspecification, i.e. how the limit θ^* of a GMC $\hat{\theta}$ behaves as the model $\mathcal{P}$ moves away from the true probability measure μ. The other is the dispersion behavior of the estimator, such as its asymptotic variance. (Some claims on this issue can be found in Schennach (2004).) As far as one considers a global misspecification (as opposed to a local misspecification, in which the model approaches the true probability measure at a certain rate), the former is typically dominant of the two, which makes the latter a second order issue. An alternative approach to the robustness issue is to consider the effect of local misspecification within a shrinking topological neighborhood of the true probability distribution, so that both bias and variance matter asymptotically. Such analysis, put loosely, enables the researcher to compare robustness in terms of MSE. This approach appears to be useful if one is interested in analyzing the robustness of empirical likelihood and other methods. Note that this line of research has been carried out in the robustness literature on parametric models. Some researchers in this literature argue for the use of minimum Hellinger distance methods. This is interesting because the GEL/GMC families include the Hellinger distance as a special case, since it is the Cressie-Read divergence with $\alpha = -\frac{1}{2}$. Detailed investigation toward a robustness theory of empirical likelihood is left for a separate paper.

8. Computational issues and numerical examples

Empirical likelihood or its generalizations have desirable theoretical properties, as described in the foregoing sections. This section turns to practical matters. First, issues associated with actual implementation are explored. Some numerical algorithms are discussed. Second, numerical examples of some of the methods discussed in the preceding sections are presented.

8.1. **Implementing Empirical Likelihood.** Computational issues for empirical likelihood are best described by considering the unconditional moment restrictions model (2.2). It appears that the most stable way to compute the EL estimator $\hat{\theta}_{\mathrm{EL}}$ is to utilize the profile likelihood at θ as given by (2.5) and write a nested optimization routine. See Chapter 12 in Owen (2001) for this and other types of algorithms. The nested optimization method requires a routine for the *inner loop minimization*, which takes θ as an argument and return the value

$$\min_{\gamma \in \mathbb{R}^q} Q_n(\theta, \gamma), \quad Q_n(\theta, \gamma) = -\sum_{i=1}^n \log(1 + \gamma' g(z_i, \theta)). \tag{8.1}$$

This is equal to the profile likelihood function $\ell(\theta)$ in (2.5), up to a constant which is irrelevant in estimating θ. Once this routine is defined, it is maximized with respect to θ. This part can be called the *outer loop maximization.* To compute $\hat{\theta}_{\mathrm{EL}}$, one uses a nested optimization algorithm where the outer maximization loop encloses the inner minimization loop. Some comments on the inner loop and the outer loop are in order. In particular, the problem associated with the situation where the convex hull spanned by $g(z_i, \theta), i = 1, .., n$ does not include the origin deserves a special attention.

The objective function Q_n in the inner loop is convex in γ. Moreover, the analytical expressions for its Jacobian and Hessian are readily available:

$$\nabla_\gamma Q_n(\theta, \gamma) = -\sum_{i=1}^n \frac{g(z_i, \theta)}{1 + \gamma' g(z_i, \theta)}, \nabla_{\gamma\gamma} Q_n(\theta, \gamma) = \sum_{i=1}^n \frac{g(z_i, \theta) g(z_i, \theta)'}{(1 + \gamma' g(z_i, \theta))^2}.$$

It is therefore reasonable to carry out Newton iterations using these expressions. Hansen (2006) suggests a rule for the choice of Newton step lengths. Even though the Hessian is positive definite by its definition, sometimes inverting it numerically is difficult when the model is not well-behaved, in particular for a value of θ that is distant from the optimal point $\hat{\theta}_{\mathrm{EL}}$. In such a situation one may consider replacing it according to a quasi-Newton method or further modifying it along the line suggested by Shanno (1970). Alternatively, one can use a nonlinear numerical optimization routine based on numerical derivatives to minimize Q_n with respect to γ.

Sometimes it may be advantageous to transform the parameter space of γ using (an approximation for) the Hessian matrix, as often done in a nonlinear numerical optimization algorithm. For the inner loop problem above, this can be achieved by premultiplying $g(z_i, \theta)$ with an appropriate matrix, such as the inverse of the Cholesky decomposition of $\sum_{i=1}^{n} g(z_i, \theta)g(z_i, \theta)'$, which is suggested by Bruce Hansen in his GAUSS code for EL. This does not change the value function $\min_{\gamma \in \mathbb{R}^q} Q_n(\theta, \gamma)$, but the definition of γ changes to $\gamma^* = [\sum_{i=1}^{n} g(z_i, \theta)g(z_i, \theta)']^{1/2'}\gamma$. The coordinate change is likely to make the Hessian $\nabla_{\gamma^*\gamma^*}Q$ close to the q dimensional identity matrix, and it may help the convergence of the optimization process when the dimension q is high.

The inner loop optimization is generally a well-behaved convex programming problem, when there is a solution. In some situations, however, it does not have a solution. This should be clear from the primal problem (2.3). If for a given θ the condition

$$\textbf{(C)} \qquad 0 \in \mathrm{co}\{g(z_1, \theta), ..., g(z_n, \theta)\}$$

fails to hold, that is, the convex hull of the n vectors of the moment function evaluated at the observations $\{z_i\}_{i=1}^{n}$ does not include the origin of $\mathbb{R}^q$, the problem (2.3) does not have a feasible solution. Note that this is more of a practical problem than a theoretical one. If (**C**) fails, it is theoretically appropriate to set the value of the empirical likelihood $\ell(\theta)$ at $-\infty$ as a convention. After all, a failure of (**C**) at a value of θ should be regarded as strong evidence against the possibility that it is the true value. As far as $E[g(z, \theta_0)] = 0$, which holds if the model is correctly specified, the condition (**C**) holds with probability approaching one at $\theta = \theta_0$.

In practice, however, (**C**) can fail in finite samples even if the model is correctly specified, partly because a numerical search in the outer maximization loop can take θ to areas that are far from the true value θ_0. Also, the vectors $\{g(z_i, \theta)\}_{i=1}^{n}$ are more likely to fail to span the origin, if the dimension of the space becomes higher or the number of the vectors becomes smaller. In other words, the condition (**C**) may occasionally fail for a large q and/or a small n. Finally, when the model is misspecified as discussed in Section 7, this becomes an important issue. One needs to proceed with caution when it happens.

Consider again the inner loop minimization problem (8.1), and suppose one starts a Newton algorithm from a initial value (e.g. $\gamma = 0$). If (**C**) fails, a Newton iteration would make γ grow (in absolute value). Theoretically, γ that "solves" the inner loop minimization (8.1) should be at infinity in a direction where $\gamma' g(z_i, \theta)$ is positive for all i.[6] This makes the value of Q_n negative infinity, which is consistent with the convention introduced above. It is, however, likely to cause a problem when implementing a numerical algorithm such as Newton's method. For example, the first-order gradient can still be large at the end of the algorithm if the maximum number of iterations is set too low. More importantly, when the elements of γ are large in absolute value, it is likely that some of the logs in Q_n would have negative arguments at a γ value the (Newton) algorithm "tries." This causes a naive algorithm to stop. It is the author's impression that a common mistake is to use a rather arbitrary value such as zero to impute the value of $\min_{\gamma \in \mathbb{R}^q} Q_n(\theta, \gamma)$ when an algorithm halts for this situation. Recall that theoretically the value of empirical log-likelihood should be negative infinity in the event of the failure of (**C**), so using inappropriate values for this situation leads to quite misleading results. This might explain some puzzling results of Monte Carlo experiments reported in the literature. This is especially relevant when the power of ELR or the behavior of the EL estimator under misspecification are being evaluated by simulations, since obviously the failure of (**C**) is an issue in these situations. It can be in principle prevented by assigning an appropriate value when this happens, but this should be done with caution as well, so that the numerical search in the outer maximization loop over θ would not remain trapped in the region where the violation of (**C**) occurs.

A practical approach to deal with the above problem is to modify the algorithm to prevent the problem associated with potential negative values in log terms of Q_n. One possibility is to use a constrained optimization routine to optimize Q_n while keeping the arguments of the log terms positive. That is, the objective function $Q_n(\theta, \gamma)$ for a given value of θ is minimized over the region $\{\gamma \in \mathbb{R}^q : 1 + \gamma' g(z_i, \theta) \geq \delta \text{ for all } i\}$ in the inner loop, where δ is a small number chosen by the econometrician. The resulting minimum values of Q_n is then maximized over θ in the outer loop. This method appears to work reasonably well in practice, even in a situation where the problem associated with the violation of (**C**) is rather severe. Another potential solution, suggested by Owen (2001), is to replace log in Q_n by a function that allows negative arguments. He suggests choosing a

[6] The separating hyperplane theorem shows that such a direction exists if the condition (**C**) is violated.

small number $\delta > 0$ and use

$$\log_\star(y) = \begin{cases} \log(y) & \text{if} \quad y > \delta \\ \log(y) - 1.5 + 2\frac{y}{\delta} - \frac{z^2}{2\delta^2} & \text{if} \quad y \leq \delta, \end{cases}$$

which is twice continuously differentiable and concave. This makes the objective function

$$Q_{\star n}(\theta, \gamma) = -\sum_{i=1}^{n} \log_\star(1 + \gamma' g(z_i, \theta))$$

well-defined for all $\gamma \in \mathbb{R}^q$.

Once the empirical likelihood function is calculated, it can be used for inference as seen in the preceding sections. Standard empirical likelihood ratio statistics possess χ^2 limiting distributions and therefore provide methods for asymptotically valid inference. Moreover, under regularity conditions, the second order derivative of the empirical log likelihood function $\ell(\theta)$ normalized by $-\frac{1}{n}$ and evaluated at $\hat{\theta}_{\text{EL}}$ converges to the appropriate asymptotic variance matrix of $\hat{\theta}_{\text{EL}}$, i.e.

$$-\frac{1}{n}\nabla_{\theta\theta}\ell(\hat{\theta}_{\text{EL}}) \xrightarrow{p} (D'SD)^{-1},$$

This can be used to obtain asymptotic standard error estimates, though the optimal power results (e.g. the generalized Neyman-Pearson property described in Section 4) strongly indicate that the empirical likelihood ratio test has theoretical advantages over other methods, including a Wald-type test or an LM-type test based on the asymptotic covariance matrix estimate $-\frac{1}{n}\nabla_{\theta\theta}\ell(\hat{\theta}_{\text{EL}})$. The same optimality results also imply that it is best to invert the likelihood ratio test statistic to obtain a confidence interval if it is computationally feasible. Many simulation studies report that empirical likelihood ratio based confidence intervals tend to be substantially shorter than other asymptotically valid intervals with comparable coverage probabilities.

In terms of size, however, the asymptotic chi-square approximation of the empirical likelihood ratio statistic may not be accurate enough when the sample size is small. One potential way to correct this is to use Bartlett adjustment discussed in Section 5.2. Analytical expressions for Bartlett factors tend to be complicated even for a relatively simple model, and are probably hard to derive for a complex structural model often used in econometrics. An alternative way to improve the size properties of empirical likelihood ratio tests is to apply the bootstrap, sometimes called the bootstrap calibration in the literature. Consider again the problem of testing the hypothesis $R(\theta_0) = 0$ discussed in Section 2, where the empirical likelihood ratio statistic r has been introduced. The nonparametric bootstrap can be implemented as follows. Resample $\{z_i\}_{i=1}^n$ according to the empirical measure μ_n

with replacements in the usual manner to obtain bootstrapped data $\{z_i^{*(b)}\}_{i=1}^n$, $b = 1, ..., B$, where B is the number of bootstrap replications. Define

$$\ell^{*(b)}(\theta) = \min_{\gamma \in \mathbb{R}^q} - \sum_{i=1}^n \log(1 + \gamma' g(z_i^{*(b)}, \theta)) - n \log n, \quad b = 1, ..., B.$$

The $b-$th bootstrap version of the empirical likelihood ratio statistic is

$$r^{*(b)} = -2 \left(\sup_{\theta \in \Theta : R(\theta) = R(\hat{\theta}_{\mathrm{EL}})} \ell^{*(b)}(\theta) - \sup_{\theta \in \Theta} \ell^{*(b)}(\theta) \right), b = 1, ..., B.$$

Alternatively, one may resample according to an EL-based estimate of the probability measure, borrowing the idea of "efficient bootstrapping" by Brown and Newey (2002). To apply their method to the current problem, define the constrained EL estimator $\hat{\theta}^c_{\mathrm{EL}} = \mathrm{argmax}_{\theta \in \Theta : R(\theta) = 0}\, \ell(\theta)$ and calculate the NPMLE weights as in (2.6) under the constraint $R(\theta) = 0$:

$$\hat{p}^c_{\mathrm{EL}i} = \frac{1}{n(1 + \hat{\gamma}(\hat{\theta}^c_{\mathrm{EL}})' g(z_i, \hat{\theta}^c_{\mathrm{EL}})}, \quad i = 1, ..., n.$$

One would then resample $\{z_i\}_{i=1}^n$ according to the probability measure $\hat{\mu}^c = \sum_{i=1}^n \hat{p}^c_{\mathrm{EL}i} \delta_{z_i}$ to generate $\{z_i^{*(b)}\}_{i=1}^n$, $b = 1, ..., B$, from which bootstrap empirical likelihood ratio statistics are obtained:

$$r^{*(b)} = -2 \left(\sup_{\theta \in \Theta : R(\theta) = 0} \ell^{*(b)}(\theta) - \sup_{\theta \in \Theta} \ell^{*(b)}(\theta) \right), b = 1, ..., B.$$

The $(1 - \alpha)-$quantile of the distribution of $\{r^{*(b)}\}_{b=1}^B$ can be then used as a bootstrap $100(1 - \alpha)\%$ critical value for r.

It is also possible to bootstrap the empirical likelihood ratio test statistic $\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})$ for overidentifying restrictions in similar ways. One way is to use the empirical distribution μ_n for resampling to generate $\{z_i^{*(b)}\}_{i=1}^n$, $b = 1, ..., B$, then replace $\{\{g(z_i^{*(b)}, \theta)\}_{i=1}^n\}_{b=1}^B$ by

$$\tilde{g}(z_i^{*(b)}, \theta) = g(z_i^{*(b)}, \theta) - \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}_{\mathrm{EL}}), \quad i = 1, ..., n, b = 1, ..., B.$$

This replacement is introduced to deal with the issue associated with bootstrapping under overidentifying restrictions (see, for example, Hall and Horowitz (1996)). Now compute

$$\mathrm{elr}^{*(b)}(\theta) = \max_{\gamma \in \mathbb{R}^q} 2 \sum_{i=1}^n \log(1 + \gamma' \tilde{g}(z_i^{*(b)}, \theta)), \quad b = 1, ..., B.$$

Evaluate each $\mathrm{elr}^{*(b)}(\theta)$ at its maximizer $\hat{\theta}^{*(b)}$ to obtain $\mathrm{elr}^{*(b)}(\hat{\theta}^{*(b)}_{\mathrm{EL}})$. The bootstrap empirical distribution of these values yields critical values for $\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})$. If, however, one uses $\hat{\mu}_{\mathrm{EL}} = \sum_{=1}^n \hat{p}_{\mathrm{EL}} \delta_{z_i}$ derived in Section 2 for resampling as in Brown and Newey (2002), the recentering step is unnecessary.

The above discussions focused on the bootstrap calibration in which the researcher uses the distribution of bootstrap versions of empirical likelihood ratio statistics in place of the appropriate chi-square distribution. There is another potentially interesting way to use bootstrap test statistic values to improve accuracy of inference. Recall that empirical likelihood ratio statistics are generally Bartlett-correctable (Section 5.2). The essence of Bartlett correction is to adjust a likelihood ratio statistic (parametric or empirical) by its expected value. Suppose, for example, one wishes to compare $\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})$ with its limiting distribution χ^2_{q-k}. Then $E[\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})] = (q-k)(1+n^{-1}a) + O(n^{-2})$ for the Bartlett factor a, and the Bartlett-corrected statistic is $\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})/(1+n^{-1}a)$. The factor a needs to be estimated, but its expression can be overwhelmingly complex. One can, however, estimate the factor in the denominator by taking the average of bootstrapped statistics generated by either of the two algorithms described above. This yields a Bartlett-corrected empirical likelihood ratio statistic via bootstrapping:

$$\frac{(q-k)\mathrm{elr}(\hat{\theta}_{\mathrm{EL}})}{\frac{1}{B}\sum_{b=1}^{B}\mathrm{elr}^{*(b)}(\hat{\theta}^{*(b)}_{\mathrm{EL}})}$$

The distribution of the above statistic can be approximated by the χ^2_{q-k} distribution up to errors of order $O(n^{-2})$.

Bootstrap Bartlett correction has been used in parametric likelihood ratio testing. In particular, Rocke (1989) considers the parametric LR test in a seemingly unrelated regression model and finds that bootstrap Bartlett correction achieves accuracy comparable to conventional bootstrap methods with a substantially smaller number of bootstrap replications. See also Zaman (1996) on the topic. It is therefore worthwhile to consider the use of bootstrap Bartlett correction for empirical likelihood in complex models where bootstrapping is costly. Chen, Leung, and Qin (2003) report a striking performance of the bootstrap Bartlett correction for their empirical likelihood ratio test with validation data.

8.2. **Simulation Results.** The theoretical analysis in Sections 4 and 5 indicates that empirical likelihood-based methods possess theoretical advantages over other competing methods. The following numerical examples provide some insights on finite sample properties of these estimators.

8.2.1. *Experiment 1.* The first simulation design is taken from Blundell and Bond (1998) and Bond, Bowsher, and Windmeijer (2001). The focus of this experiment is the relative finite sample performance of EL-based estimators and the conventional GMM.[7] It is concerned with a dynamic panel

[7] Further details of this experiment are to be found in Kitamura and Otsu (2005).

data model: $y_{it} = \theta_0 y_{t-1} + \eta_i + u_{it}, i = 2, ..., n, t = 1, ..., T$ where $\eta_i \sim_{iid} \mathrm{N}(0,1)$, $u_{it} \sim_{iid} \mathrm{N}(0,1)$, and $e_i \sim_{iid} \mathrm{N}(0, \frac{1}{1-\theta_0^2})$, and these shocks are independent. The initial value is drawn according to $y_{i1} = \frac{\eta_i}{1-\theta_0} + e_i$. The two equations, together with the independence assumptions, imply that

$$E[y_{i,s}(\Delta y_{it} - \theta_0 \Delta y_{it-1})] = 0, t = 1, ..., T, s = 1, ..., t-2 \tag{8.2}$$

and

$$E[\Delta y_{it-1}(y_{it} - \theta_0 y_{it-1})] = 0, t = 3, ..., T. \tag{8.3}$$

The task is to estimate the parameter θ_0 using the moment conditions (8.2) and (8.3). The following estimators are considered: (i) the 2-step GMM with its weighting matrix obtained by the usual robust estimator as described in Blundell and Bond (1998) ($\hat{\theta}_{\mathrm{GMM}}$), (ii) the continuous updating GMM by Hansen, Heaton, and Yaron (1996) ($\hat{\theta}_{\mathrm{CUE}}$), (iii) the maximum empirical likelihood estimator ($\hat{\theta}_{\mathrm{EL}}$), and (iv) and the minimax estimator by Kitamura and Otsu (2005) ($\hat{\theta}_{\mathrm{ld}}$) with $c = 0.1$ and 0.2. The second design is based on Bond, Bowsher, and Windmeijer (2001), where u_{it} is replaced by a conditionally heteroskedastic process of the form $u_{it}|y_{it-1} \sim \mathrm{N}(0, 0.4 + 0.3y_{it-1}^2)$. The initial condition is generated using fifty pre-sample draws as in Bond, Bowsher, and Windmeijer (2001). The third design is the same as the first, except that u_{it} is the (standardized) chi-square distribution with one degree of freedom: $u_{it} \sim_{iid} (\chi_1^2 - 1)/\sqrt{2}$. Experimenting with asymmetric errors such as this specification is important, since one of the main advantages of GMM, EL or other moment-based estimators is its robustness against distributional assumptions. Also, asymmetric shocks appear to be an important characteristic of empirical models of income dynamics (see Geweke and Keane (2000) and Hirano (2002)), which is one of the main applications of dynamic panel data models. For these three designs, the true value for the autoregressive parameter θ is set at 0.9. The fourth is the same as the first, except that the AR parameter θ is set at 0.4. The panel dimensions are $n = 100$ and $T = 6$, and the number of Monte Carlo replications is 1000 for each design. The first and second panels of Table 1 (Table 2) display results from the first and second (the third and the fourth) designs, respectively. The five columns of each panel correspond to bias, root mean square errors (RMSE), mean absolute errors (MAE) and the probabilities of the estimators deviating from the true value by more than $d = 0.1$ and 0.2, respectively.

The results of the experiment are intriguing, and in accordance with the theoretical results presented in Sections 4 and 5. Some caution needs to be exercised in interpreting figures such as

RMSE, as the existence of moments of these estimators can be an issue; see, e.g. Kunitomo and Matsushita (2003).

In the first design with homoskedastic and normal idiosyncratic shocks, all of the estimators work reasonably well, though the minimax estimation method by Kitamura and Otsu (2005) leads to substantial efficiency gain in terms of MAE. Also, the deviation probabilities for $d = .1$ are much lower for the minimax estimators than for CUE and EL. While the performance of (two-step) GMM is only slightly worse than that of the minimax estimators in this design, that changes dramatically in the second design, where conditional heteroskedasticity is introduced. Even though the bias of the minimax estimators is slightly inflated relative to that of EL, it is offset by their variance reduction. This is consistent with our interpretation that the minimax method "robustifies" the original EL estimator. Deviation probabilities for this design exhibit an interesting pattern. Take the case with $d = 0.2$. GMM falls outside of the interval 0.8 ± 0.2 with 61 percent probability. CUE is much better

Table 1. Estimation of Dynamic Panel Data Model (1)

		homoskedastic u_{it}					heteroskedastic u_{it}				
		bias	RMSE	MAE	$\Pr\{\lvert\theta_n - \theta_0\rvert > d\}$		bias	RMSE	MAE	$\Pr\{\lvert\theta_n - \theta_0\rvert > d\}$	
					$d = .1$	$d = .2$				$d = .1$	$d = .2$
$\hat{\theta}_{\text{GMM}}$		.014	.096	.071	.296	.029	-.253	.364	.261	.815	.614
$\hat{\theta}_{\text{CUE}}$		.001	.113	.084	.390	.054	-.080	.264	.148	.643	.368
$\hat{\theta}_{\text{EL}}$		-.005	.113	.080	.370	.056	-.059	.189	.119	.570	.275
$\hat{\theta}_{\text{ld}}$	c=.1	-.016	.100	.061	.274	.047	-.064	.182	.110	.542	.258
	c=.2	-.027	.090	.056	.233	.037	-.076	.166	.100	.503	.215

Table 2. Estimation of Dynamic Panel Data Model (2)

		homoskedastic & asymmetric u_{it}					$\theta = 0.4$, homoskedastic u_{it}				
		bias	RMSE	MAE	$\Pr\{\lvert\theta_n - \theta_0\rvert > d\}$		bias	RMSE	MAE	$\Pr\{\lvert\theta_n - \theta_0\rvert > d\}$	
					$d = .1$	$d = .2$				$d = .1$	$d = .2$
$\hat{\theta}_{\text{GMM}}$		-0.221	.312	.230	.804	.580	-.005	.134	.091	.457	.124
$\hat{\theta}_{\text{CUE}}$		.002	.213	.168	.732	.404	-.025	.141	.095	.477	.131
$\hat{\theta}_{\text{EL}}$		-.023	.176	.137	.627	.306	-.0018	.119	.079	.388	.075
$\hat{\theta}_{\text{ld}}$	c=.1	-.022	.162	.125	.575	.247	-.0016	.115	.076	.373	.067
	c=.2	-.029	.134	.093	.472	.141	-.0010	.104	.070	.340	.053

than GMM (37 percent), though still it is high. EL has a good performance (28 percent), and by using the minimax method with $c = 0.2$, the probability is reduced to 21 percent, nearly one third of that of GMM. The last result seems to support the theoretical results by Kitamura and Otsu (2005) discussed in Section 4.1. Similar patterns emerge for the third design with asymmetric errors; again, note the drastic reduction of the deviation probability with $d = .2$ by the minimax estimator with $c = .2$ (the probability is 14 percent, compared with the 58 deviation probability of GMM, for example). In the fourth design, EL and its minimax versions continue to outperform GMM and CUE.

8.2.2. *Experiment 2.* This experiment is concerned with testing, in particular the power properties of the empirical likelihood ratio test. Pseudo-samples $\{z_i\}_{i=1}^n$ are independently drawn from a distribution F, and by adding a location shift term c, $x_i = z_i + c, i = 1, ..., n$, are calculated. Three specifications of F are considered: (1) standard normal $\Phi(z)$, (2) normal mixture $.1\Phi(z-9) + .9\Phi(z+1)$, and (3) lognormal $\Phi(\log(z)), z > 0$. The null hypothesis is: $E[x] = 0$. This is the simplest possible example of overidentifying restrictions: the number of moment conditions is one and no parameters are estimated, so the degree of overidentification is one. Two statistics are used to test this null; one is the empirical likelihood ratio ℓ_{EL} and the other is $W = n(\bar{x}^2)/n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$, where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$ ("W" stands for "Wald"). The "Wald" statistic is a feasible J-statistic in this simple setting. The sample size n is set to be 50. Tables 3-6 report rejection frequencies of the two tests.

The standard normal distribution belongs to the family of distributions discussed by Kariya (1981), for which W is Uniformly Most Powerful Invariant (UMPI). In this sense, the experimental design with normal z's (Table 3) is favorable to W, since no other invariant test should outperform W after size correction, for any finite sample size. Nevertheless, Table 3 shows that the power of

TABLE 3. Standard Normal, Size = 0.01

	Size Uncorrected		Size Corrected	
c	ℓ_{EL}	W	ℓ_{EL}	W
0.0	0.012	0.013	0.010	0.010
0.3	0.322	0.348	0.300	0.303
0.5	0.809	0.832	0.789	0.794

TABLE 4. Normal Mixture, Size = 0.01

	Size Uncorrected		Size Corrected	
c	ℓ_{EL}	W	ℓ_{EL}	W
-1.2	0.887	0.574	0.868	0.028
-0.6	0.174	0.043	0.148	0.001
0.0	0.011	0.041	0.010	0.010
0.6	0.082	0.206	0.073	0.075
-1.2	0.344	0.553	0.320	0.263

Table 5. Normal Mixture, Size = 0.05

	Size Uncorrected		Size Corrected	
c	ℓ_{EL}	W	ℓ_{EL}	W
-1.2	0.961	0.876	0.960	0.729
-0.6	0.361	0.199	0.353	0.093
0.0	0.055	0.085	0.050	0.050
0.6	0.225	0.348	0.207	0.224
-1.2	0.614	0.727	0.594	0.605

Table 6. Lognormal, Size = 0.01

	Size Uncorrected		Size Corrected	
c	ℓ_{EL}	W	ℓ_{EL}	W
-1.0	0.582	0.752	0.404	0.468
-0.6	0.325	0.480	0.176	0.201
0.0	0.034	0.056	0.010	0.010
0.6	0.640	0.248	0.421	0.003
1.0	1.000	0.947	0.998	0.338

the empirical likelihood ratio keeps up with that of W reasonably well. (The power curves for the standard normal are symmetric, so only the results for nonnegative c's are reported in Table 3.)

In the normal mixture and lognormal cases, the size distortion of W makes power comparison difficult and misleading, and size corrected power might give a better picture. Table 4 shows the excellent power properties of the empirical likelihood ratio test. When the deviation from the null is $c = -1.2$, the power of ℓ_{EL} is nearly 90 percent, whereas the power of W is extremely poor (2.8 percent). Qualitatively similar results are obtained for larger nominal sizes (see Table 5), and for other distributions such as the lognormal (see Table 6). In summary, the simulation results seem to be consistent with the large deviation optimality results of the empirical likelihood ratio test in Section 4.

9. Conclusion

This paper has discussed several aspects of empirical likelihood. Two different but interconnected interpretations for empirical likelihood have been offered. One can view empirical likelihood as NPMLE, which has a long history in statistics. The literature on empirical likelihood initiated by Owen (1988) demonstrates that NPMLE applied to a moment restriction model yields an attractive procedure, both practically and theoretically. Moreover, applications of empirical likelihood extend to other problems that are important in applied economics, as discussed in the present paper. Alternatively, one can view empirical likelihood as GMC with a particular choice of the "contrast function." This line of argument yields a variety of empirical likelihood-type estimators and tests, depending on the choice of the contrast function. The theory of convex duality shows a clear connection between GMC and other related estimators, including Smith's GEL. Theoretical considerations seem to indicate that the contrast function used for empirical likelihood is often the most preferred choice.

A natural conjecture sometimes made in the literature is that empirical likelihood may bring efficiency properties analogous to those of parametric likelihood to semiparametric analysis, while retaining the distribution-free properties of certain nonparametric procedures. The results described in this paper present affirmative answers to this conjecture. In particular, the large deviation principle (LDP) provides compelling theoretical foundations for the use of empirical likelihood through Sanov's theorem.

Another attractive aspect of empirical likelihood is that it directly uses the empirical distribution of the data, which has intuitive and practical appeal. It avoids, or at least lessens, the problem of choosing tuning parameters that often introduce a fair amount of arbitrariness to nonparametric and semiparametric procedures. A related and important point is the practicality of empirical likelihood. The use of convex duality transforms seemingly complex optimization problems into their simple dual forms, thereby making empirical likelihood a highly usable method. This paper has provided discussions on the implementation of empirical likelihood as well as numerical examples, so that they offer practical guidance to applied economists who wish to use empirical likelihood in their research.

10. Appendix

Derivation of Equations (5.3) and (5.4). The objective function to be maximized is

$$-\sum_{i=1}^{n}\log(1+\hat{\gamma}(\theta)'g(z_i,\theta)).$$

Consider the first order condition. Since the $\hat{\gamma}(\theta)$ is an optimizer (for a given θ), the derivative of $\hat{\gamma}$ with respect to θ drops out by the envelop theorem. Therefore:

$$\sum_{i=1}^{n}\frac{\nabla_\theta g(z_i,\hat{\theta}_{\mathrm{EL}})'\hat{\gamma}}{1+\hat{\gamma}'g(z_i,\hat{\theta}_{\mathrm{EL}})}=0,\quad \hat{\gamma}=\hat{\gamma}(\hat{\theta}_{\mathrm{EL}}). \tag{10.1}$$

Now, the first order condition for $\hat{\gamma}$ is

$$\sum_{i=1}^{n}\frac{g(z_i,\theta)}{1+\hat{\gamma}'g(z_i,\theta)}=0.$$

Manipulating this yields

$$\hat{\gamma}=\left[\sum_{i=1}^{n}\frac{g(z_i,\hat{\theta}_{\mathrm{EL}})g(z_i,\hat{\theta}_{\mathrm{EL}})'}{n(1+\hat{\gamma}'g(z_i,\hat{\theta}_{\mathrm{EL}}))}\right]^{-1}\bar{g}(\hat{\theta}_{\mathrm{EL}}). \tag{10.2}$$

By (10.1) and (10.2),

$$\left[\sum_{i=1}^{n}\frac{\nabla_\theta g(z_i,\hat{\theta}_{\mathrm{EL}})}{1+\hat{\gamma}'g(z_i,\hat{\theta}_{\mathrm{EL}})}\right]'\left[\sum_{i=1}^{n}\frac{g(z_i,\hat{\theta}_{\mathrm{EL}})g(z_i,\hat{\theta}_{\mathrm{EL}})'}{n(1+\hat{\gamma}'g(z_i,\hat{\theta}_{\mathrm{EL}}))}\right]^{-1}\bar{g}(\hat{\theta}_{\mathrm{EL}})=0.$$

Use the definition of $\hat{p}_{\mathrm{EL}i}$ given by (2.6) to obtain (5.3).

To obtain (5.4), differentiate $\bar{g}(\theta)'\bar{S}(\theta)^{-1}\bar{g}(\theta)$ by θ (assume that θ is a scalar for the ease of presentation) to obtain

$$\begin{aligned}
&\nabla_\theta \bar{g}(\hat{\theta}_{\mathrm{cue}})'\bar{S}^{-1}(\hat{\theta}_{\mathrm{cue}})\bar{g}(\hat{\theta}_{\mathrm{cue}}) - \bar{g}(\hat{\theta}_{\mathrm{cue}})\bar{S}^{-1}(\hat{\theta}_{\mathrm{cue}})\frac{1}{n}\sum_{i=1}^{n} g(z_i,\hat{\theta}_{\mathrm{cue}})\nabla_\theta g(z_i,\hat{\theta}_{\mathrm{cue}})'\bar{S}^{-1}(\hat{\theta}_{\mathrm{cue}})\bar{g}(\hat{\theta}_{\mathrm{cue}})\\
&\quad= \left[\nabla_\theta \bar{g}(\hat{\theta}_{\mathrm{cue}}) - \left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta g(z_i,\hat{\theta}_{\mathrm{cue}})g(z_i,\hat{\theta}_{\mathrm{cue}})\right)\bar{S}^{-1}(\hat{\theta}_{\mathrm{cue}})\bar{g}(\hat{\theta}_{\mathrm{cue}})\right]'\bar{S}^{-1}(\hat{\theta}_{\mathrm{cue}})\bar{g}(\hat{\theta}_{\mathrm{cue}})\\
&\quad= \tilde{D}(\hat{\theta}_{\mathrm{cue}})'\bar{S}^{-1}(\hat{\theta}_{\mathrm{cue}})\bar{g}(\hat{\theta}_{\mathrm{cue}})\\
&\quad= 0.
\end{aligned}$$

□

References

Aït-Sahalia, Y., P. Bickel, and T. Stoker (2001): "Goodness-of-fit tests for kernel regression with an application to option implied volatilities," *Journal of Econometrics*, 105, 363–412.

Akaike, H. (1973): "Information theory and an extension of the likelihood ratio principle," in *Proceedings of the Second International Symposium of Information Theory*. Budapest: Akademiai Kiado, pp. 257–281.

Allen, J., A. Gregory, and K. Shimotsu (2005): "Empirical likelihood block bootstrap," Manuscript, Department of Economics, Queen's University.

Altonji, J., and L. M. Segal (1996): "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal Business and Economic Statistics*, 14, 353–366.

Anatolyev, S. (2005): "GMM, GEL, Serial correlation, and asymptotic bias," *Econometrica*, 73, 983–1002.

Anderson, T. W., and H. Rubin (1949): "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570–582.

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59, 817–858.

Andrews, D. W., and C. Monahan (1992): "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator," *Econometrica*, 60, 953–966.

Angrist, J., V. Chernozhukov, and I. Fernandez (2005): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structures," *Econometrica*, forthcoming.

Antoine, B., H. Bonnal, and E. Renault (2006): "On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Likelihood," *Journal of Econometrics*, forthcoming.

Baggerly, K. (1998): "Empirical likelihood as a goodness-of-fit measure," *Biometrika*, 85, 535–547.

Bahadur, R. (1960): "On the asymptotic efficiency of tests and estimators," *Sankhyā*, 22, 229–252.

——— (1964): "On Fisher's Bound for Asymptotic Variances," *Annals of Mathematical Statistics*, 35, 1545–1552.

Bahadur, R., S. Zabell, and J. Gupta (1980): "Large deviations, tests, and estimates," in *Asymptotic Theory of Statistical Tests and Estimation*, ed. by I. M. Chaterabarli. New York: Academic Press, pp. 33–64.

Bickel, P., C. Klassen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins Press.

Blundell, R., and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

Bond, S., C. Bowsher, and F. Windmeijer (2001): "Criterion-Based Inference for GMM in Autoregressive Panel Data Models," *Economics Letters*, 73, 379–388.

Borwein, J. M., and A. S. Lewis (1991): "Duality relationships for entropy-type minimization problems," *SIAM Journal of Control and Optimization*, 29, 325–338.

Bravo, F. (2002): "Blockwise empirical Cressie-Read test statistics for α-mixing processes," *Statistics and Probability Letters*, 58, 319–325.

——— (2005a): "Blockwise empirical entropy tests for time series regressions," *Journal of Time Series Analysis*, 26, 157–321.

——— (2005b): "Sieve empirical likelihood for unit root tests," Manuscript, University of York.

Brown, B. W., and W. K. Newey (2002): "Generalized method of moments, efficient bootstrapping, and improved inference," *Journal of Business and Economic Statistics*, 20, 507–517.

Brown, D. J., and M. H. Wegkamp (2002): "Weghted minimum mean-square distance from independence estimation," *Econometrica*, 70, 2035–2051.

Caner, M. (2003): "Exponential tilting with weak instruments: Estimation and testing," Manuscript, University of Pittsburgh.

Carlstein, E. (1986): "The use of subseries values for estimating the variance of a general statistic from stationary processes," *Annals of Statistics*, 14, 1171–1179.

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

Chen, S. X., and H. Cui (2004): "On the second order properties of empirical likelihood with moment conditions," Manuscript.

——— (2005): "On Bartlett correction of empirical likelihood in the presence of nuisance parameters," *Biometrika*, Forthcoming.

Chen, S. X., W. Härdle, and M. Li (2003): "An empirical likelihood goodness-of-fit test for time series," *Journal of The Royal Statistical Society, Series B*, 65, 663–678.

Chen, S. X., D. H. Y. Leung, and J. Qin (2003): "Information recovery in a study with surrogate endpoints," *Journal of the American Statistical Association*, pp. 1052–1062.

Chen, X., H. Hong, and M. Shum (2001): "Nomparametric Likelihood Selection Tests for Parametric versus Moment Condition Models," Manuscript.

Chesher, A., and R. Smith (1997): "Likelihood ratio specification tests," *Econometrica*, 65, 627–646.

Christoffersen, P., J. Hahn, and A. Inoue (2001): "Testing and Comparing Value at Risk Measures," *Journal of Empirical Finance*, 8, 325–342.

Corcoran, S., A. C. Davison, and R. H. Spady (1995): "Reliable inference from empirical likelihood," Working Paper, Nuffield College, Oxford University.

Corcoran, S. A. (1998): "Bartlett adjustment of empirical discrepancy statistics," *Biometrika*, 85, 967–972.

Cosslett, S. (1983): "Distribution-free maximum likelihood estimator of the binary choice model," *Econometrica*, 51, 765–782.

——— (1997): "Nonparametric maximum likelihood methods," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier Science, pp. 385–404.

Csiszàr (1967): "On topological properties of f-divergences," *Studia Scientriarum Mathematicarum Hungaria*, 2, 329–339.

Dahlhaus, R., and D. Janas (1996): "A frequency domain bootstrap for ratio statistics in time series analysis," *The Annals of Statistics*, 24(5), 1934–1963.

Deuschel, J. D., and D. W. Stroock (1989): *Large Deviations*. Academic Press.

DiCiccio, T., P. Hall, and J. Romano (1991): "Empirical Likelihood is Bartlett-Correctable," *Annals of Statisics*, 19, 1053–1061.

Dominguez, M., and I. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72, 1601–1615.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions," *Journal of Econometrics*, 117, 55–93.

Donald, S. G., and W. K. Newey (2000): "A jackknife interpretation of the continuous updating estimator," *Economics Letters*, 67, 239–244.

Eubank, R., and C. Spiegelman (1990): "Testing the Goodness of Fit of a Linear Model via Nonparametric Regression Techniques," *Journal of the American Statistical Association*, 85, 387–392.

Fu, J. C. (1973): "On a theorem of Bahadur on the rate of convergence of point estimators," *Annals of Statistics*, 1, 745–749.

Gagliardini, P., C. Gourieroux, and E. Renault (2004): "Efficient Derivative Pricing by Extended Method of Moments," Working Paper.

Gali, J., and M. Gertler (1999): "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, 44, 195–222.

Geweke, J., and M. Keane (2000): "An empirical analysis of earnings dynamics among men in the PSID: 19681989," *Journal of Econometrics*, 96, 293–356.

Ghosh, J. K. (1994): *Higher Order Asymptotics*. Institute of Mathematical Statistics.

Guggenberger, P., and R. H. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification," *Econometric Theory*, Forthcoming.

Hall, P. (1985): "Resampling a Coverage Process," *Stochastic Processes and Their Applications*, 19, 259–269.

Hall, P., and J. L. Horowitz (1996): "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," *Econometrica*, 64, 891–916.

Hall, P., and B. LaScala (1990): "Methodology and Algorithms for Empirical Likelihood," *International Statistical Review*, 58, 109–127.

Hansen, B. E. (2006): "Econometrics," Manuscript, Department of Economics, University of Wisconsin.

Hansen, L., and K. Singleton (1982): "Generalized Instrumental Variable Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

Härdle, W., and E. Mammen (1993): "Comparing Nonparametric versus Parametric Regression Fits," *Annals of Statistics*, 21, 1926–1947.

Hart, J. D. (1997): *Nonparametric Smoothing and Lack-of-fit Tests*. Springer-Verlag.

Hastie, T., and R. Tibshirani (1986): "Generalized Additive Models," *Statistical Science*, 1, 297–318.

Heckman, J. J., and B. Singer (1984): "A method of minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, 52, 271–320.

Hellerstein, J. K., and G. W. Imbens (1999): "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics*, 81, 1–14.

Hirano, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781–799.

Hoeffding, W. (1963): "Asymptotically Optimal Tests for Multinomial Distributions," *Annals of Mathematical Statistics*, 36, 369–408.

Imbens, G. W. (1997): "One-Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–383.

Imbens, G. W., and R. H. Spady (2006): "The performance of empirical likelihood and its generalizations," in *Identification and Inference for Economic Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews, and J. H. Stock. Cambridge, UK: Cambridge University Press, pp. 216–244.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Joag-Dev, K., and F. Proschan (1983): "Negative association of random variables with applications," *Annals of Statistics*, 11, 286–295.

Kallenberg, W. (1983): "On moderate deviation theory in estimation," *Annals of Statistics*, 11, 498–504.

Kariya, T. (1981): "A robustness property of Hotelling's test," *Annals of Statistics*, 9, 211–214.

Kester, A., and W. Kallenberg (1986): "Large deviations of estimators," *Annals of Statistics*, 14, 648–664.

Kim, T.-H., and H. White (2002): "Estimation, Inference, and Specification Testing for Possibly Misspecified Quantile Regression," *Advances in Econometrics*, forthcoming.

Kitamura, Y. (1996a): "Empirical Likelihood and the Bootstrap for Time Series Regressions," Working Paper, Department of Economics, University of Minnesota.

——— (1996b): "GNP-Optimal Tests for Moment Restrictions," Working Paper, Department of Economics, University of Minnesota.

——— (1997): "Empirical likelihood methods with weakly dependent processes," *Annals of Statistics*, 25, 2084–2102.

——— (1998): "Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood," Working Paper, Department of Economics, University of Wisconsin.

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

——— (2002): "A Likelihood-based Approach to the Analysis of a Class of Nested and Non-nested Models," Working Paper, Department of Economics, University of Pennsylvania.

Kitamura, Y., and T. Otsu (2005): "Minimax Estimation and Testing for Moment Condition Models via Large Deviations," Manuscript, Department of Economics, Yale University.

Kitamura, Y., and M. Stutzer (1997): "An Information Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65(4), 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn (2001): "Empirical likelihood based inference in conditional moment restriction models," Manuscript, Department of Economics, University of Wisconsin-Madison.

——— (2004): “Empirical likelihood based inference in conditional moment restriction models,” *Econometrica*, 72, 1667–1714.

Kleibergen, F. (2002): “Pivotal statistics for testing structural parameters in instrumental variables regression,” *Econometrica*, 70, 1781–1803.

Koenker, R., and G. Bassett (1978): “Regression Quantiles,” *Econometrica*, 46, 33–50.

Kunitomo, N., and Y. Matsushita (2003): “On Finite Sample Distributions of the Empirical Likelihood Estimator and the GMM Estimator,” Manuscript.

Künsch, H. R. (1989): “The jacknife and the bootstrap for general stationary observations,” *Annals of Statistics*, 17, 1217–1241.

LeBlanc, M., and J. Crowley (1995): “Semiparametric Regression Functionals,” *Journal of the American Statistical Association*, 90(429), 95–105.

Lin, L., and R. Zhang (2001): “Blockwise empirical Euclidean likelihood for weakly dependent processes,” *Statistics and Probability Letters*, 53, 143–152.

Manski, C. F. (1983): “Closest empirical distribution estimation,” *Econometrica*, 51, 305–319.

Monti, A. C. (1997): “Empirical likelihood confidence regions in time series analysis,” *Biometrika*, 84, 395–405.

Newey, W. K. (1990): “Efficient Instrumental Variables Estimation of Nonlinear Models,” *Econometrica*, 58, 809–837.

Newey, W. K., and R. J. Smith (2004): “Higher order properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72, 219–255.

Newey, W. K., and K. D. West (1987): “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix,” *Econometrica*, 55(3), 703–708.

Newey, W. K., and F. Windmeijer (2006): “GMM with many weak moment conditions,” Working Paper, Department of Economics, MIT.

Nordman, D., and S. N. Lahiri (2004): “A frequency domain empirical likelihood for short- and long-range dependence,” Manuscript, Department of Economics, University of Iowa.

Nordman, D., P. Sibbertsen, and S. N. Lahiri (2006): “Empirical likelihood confidence intervals for the mean of a long-range dependent process,” *Econometric Theory*, forthcoming.

Otsu, T. (2006): “Generalized empirical likelihood under weak identification,” *Econometric Theory*, forthcoming.

Owen, A. (1988): “Empirical Likelihood Ratio Confidence Intervals for a Single Functional,” *Biometrika*, 75(2), 237–249.

——— (1990): “Empirical Likelihood Ratio Confidence Regions,” *The Annals of Statistics*, 18(1), 90–120.

——— (1991): “Empirical Likelihood for Linear Models,” *The Annals of Statistics*, 19(4), 1725–1747.

——— (2001): *Empirical Likelihood*. Chapman and Hall/CRC.

Priestley, M. B. (1981): *Spectral Analysis and Time Series*. New York: Academic Press.

Puhalskii, A., and V. Spokoiny (1998): “On large-deviation efficiency in statistical inference,” *Bernoulli*, 4, 203–272.

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

Ragusa, G. (2005): “Properties of Minimum Divergence Estimators,” Manuscript.

Ramalho, J. J. S., and R. J. Smith (2002): “Generalized empirical likelihood non-nested tests,” *Journal of Econometrics*, 107, 99–125.

Robinson, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

Rocke, D. M. (1989): "Bootstrap Bartlett adjustment in seemingly unrelated regression," *Journal of the American Statistical Association*, 84(406), 598–601.

Sanov, I. N. (1961): "On the Probability of Large Deviations of Random Variables," *Selected Translations in Mathematical Statistics and Probability*, I, 213–244.

Schennach, S. M. (2004): "Exponentially tilted empirical likelihood," Discussion Paper, University of Chicago.

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. John Wiley.

Shanno, D. F. (1970): "Parameter selection for modified Newton methods for function minimization," *SIAM Journal on Numerical Analysis*, 7(3), 366–372.

Smith, R. J. (1997): "Alternative semi-parametric likelihood approaches to generalized method of moments estimation," *Economic Journal*, 107, 503–519.

——— (2000): "Empirical Likelihood Estimation and Inference," in *Applications of Differential Geometry to Econometrics*, ed. by M. Salmon, and P. Marriott. Cambridge: Cambridge University Press, pp. 119–150.

——— (2003): "Efficient Information Theoretic Inference for Conditional Moment Restrictions," Working Paper, University of Cambridge.

——— (2004): "GEL criteria for moment condition models," Working Paper, University of Warwick.

——— (2005): "Local GEL Methods for Conditional Moment Restrictions," Working Paper, University of Cambridge.

Stock, J. H., and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68(5), 1055–1096.

Tripathi, G. (2005): "Moment based inference with stratified data," Working Paper, Department of Economics, University of Connecticut.

Tripathi, G., and Y. Kitamura (2003): "Testing Conditional Moment Restrictions," *Annals of Statisics*, 31, 2059–2095.

Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307–333.

Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426–482.

——— (1950): *Statistical Decision Functions*. New York: Wiley.

Wang, Q., and J. N. K. Rao (2002): "Empirical likelihood-based inference under imputation for missing response data," *Annals of Statistics*, 30, 896–924.

Whang, Y.-J. (2006): "Smoothed empirical likelihood methods for quantile regression models," *Econometric Theory*, forthcoming.

White, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.

——— (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

Wolfowitz, J. (1957): "The minimum distance method," *Annals of Methematical Statistics*, 28, 75–88.

You, J., G. Chen, and Y. Zhou (2006): "Block empirical likelihood for longitudinal partially linear regression models," *Canadian Journal of Statistics*, 34, forthcoming.

Zaman, A. (1996): *Statistical Foundations for Econometric Techniques*. Academic Press.

Zeitouni, O., and M. Gutman (1991): "On universal hypothesis testing via large deviations," *IEE Transactions on Information Theory*, 37, 285–290.

Zhang, J. (2006): "Empirical likelihood for NA series," *Statistics and Probability Letters*, 76, 153–160.

Zhang, J., and I. Gijbels (2003): "Sieve Empirical Likelihood and Extensions of the Generalized Least Squares," *Scandinavian Journal of Statistics*, 30, 1–24.

Department of Economics, Yale University, New Haven, CT-06520.

E-mail address: yuichi.kitamura@yale.edu