Empirical likelihood methods with application to econometrics

Francesco Bravo

12 November 2007

Lecture 1

Since its introduction as a nonparametric likelihood based alternative to likelihood and bootstrap methods, Owen (1988, 2001) empirical likelihood (EL) has gained increasing popularity among econometricians and statisticians.

This set of 5 lecture notes introduces the basic ideas behind EL and some of its generalisations, and illustrate them in the context of econometric models that are defined in terms of a set of moment conditions.

1 Plan

- Introduction to EL and some asymptotic properties
- Generalisations of EL
- Efficient probabilities and bootstrap
- GEL and smoothing and for weakly dependent observations
- GEL in some nonstandard conditions

2 Introduction

EL is a technique to obtain inferences on unknown parameters using nonparametric likelihood ratios. Many properties of parametric likelihood functions have nonparametric parallels. The main one is that there is a nonparametric version of Wilks (1938) famous result that the (log) likelihood ratio has an asymptotic chi-squared distribution. The nonparametric version has the clear advantage of holding under very weak conditions.

EL has a number of theoretically interesting and practically relevant properties -see Owen (2001)

1. The shape of confidence regions is data determined ,

2. When constraints are known to hold among the parameters of interest, they can be imposed numerically,

3. A Bartlett correction applies,

4. All points in confidence regions obey range restrictions: variances are nonnegative, probabilities are between $[0, 1]$, correlations are between $[-1, 1]$,

5. Transformation invariant and internal studentisation ,

6. Second-order maximinity (Bravo 2003),

7. It is optimal in the Generalized Neyman-Pearson Lemma (GNP) sense (Kitamura 2001).

The GNP optimality of EL implies that under certain conditions EL tests are uniformly more powerful in a Large Deviations (LD) sense (see Kitamura (2006) for a nice review of EL in connection with LD properties).

3 What is EL?

Consider the ordinary parametric likelihood method. Let $(z_i)_{i=1}^n$ denote a random sample from a known density $f(z,\theta)$, let $L(\theta) = \prod_{i=1}^n f(z_i,\theta)$ denote the likelihood function for θ, and let $\widehat{\theta} = \arg\max_\theta L(\theta)$ denote the maximum likelihood estimator.
Suppose that we are interested to test the hypothesis $H_0 : \theta = \theta_0$. Let

$$R(\theta_0) = L(\theta_0) / L\left(\widehat{\theta}\right)$$

denote the likelihood ratio statistic. Wilks' theorem shows that $-2\log R(\theta_0)$ converges in distribution to a chi-squared random variable.
EL replaces ordinary parametric likelihood with a particular nonparametric version. To be specific suppose that $(z_i)_{i=1}^n$ is a random sample from an unknown distribution F.
A nonparametric likelihood is defined by taking the definition of likelihood literally, i.e. as the probability that has generated the observed sample.

Let π_i denote the probability associated with the observed z_i, and let

$$L(F) = \prod_{i=1}^{n} \pi_i$$

denote the resulting nonparametric likelihood. EL considers only the π_is satisfying the following $\sum_{i=1}^{n} \pi_i = 1$, that is EL effectively uses a multinomial likelihood supported on the sample.
A Lagrangian argument shows in the absence of constraints the nonparametric maximum likelihood estimator for $L(F)$ is the empirical distribution function

$$\widetilde{F}(z) = \sum_{i=1}^{n} I\{z_i \leq z\}/n$$

with probability $\widetilde{\pi}_i = 1/n$.
Thus in analogy to ordinary likelihood methods we can define a nonparametric likelihood ratio as

$$R(F)^{1} = \prod_{i=1}^{n} n\pi_i.$$

[1] Let $T(F)$ denote a statistical functional. It is natural to wonder how the set $C = \{T(F) | R(F) \geq r\}$ can be used as confidence regions for $T(F_0)$, or equivalently whether tests of $T(F_0) = t$ can be constructed by rejecting if and only if $t \neq C$. It is clear that some care needs to be taken to define C otherwise $C = \mathbb{R}^k$ whenever $r < 1$. As argued by Owen (1990) this problem can be solved by assuming that the data belongs to a bounded set. The convex hull of the data suffices.

This is what EL does in practice.

4 A bit of history

The idea of nonparametric likelihood estimation has been used in statistics in particular in connection with data that are indirectly sampled or incompletely observed.
Well-known examples are for right censored data Kaplan & Meier (1958) and biased sampling Vardi (1982).
Owen (2001) notes that first use of nonparametric likelihood ratios appears to be due to Thomas & Grunkemeier (1975). They consider the survivor function $S(z)$ and use the Kaplan-Meier estimator $\widehat{S}(z)$ to define the ratio

$$R(S) = L(S) / L\left(\widehat{S}\right).$$

5 Moment conditions models and Z estimators

Let $(z_i)_{i=1}^n$ be i.i.d. observations of the data vector z from an unknown distribution F. A great deal of econometric models can be expressed in terms of a finite set of (unconditional) moment conditions of the form

$$E\left[g\left(z_i, \theta_0\right)\right] = 0, \tag{1}$$

for a unique unknown θ_0, where the expectation E is with respect F.
Assume that $\dim\left(g\left(\cdot\right)\right) = \dim\left(\theta\right)$ (i.e. the model is exactly identified).

Example 1 (Linear regression) Let $z_i = \left[y_i, x_i'\right]'$*, and let* $y_i = x_i'\theta_0 + \varepsilon_i$ *where* ε_i *is an unobservable error with* $E\left(x_i\varepsilon_i\right) = 0$*. Then*

$$E\left[x_i\left(y_i - x_i'\theta_0\right)\right] = 0.$$

Example 2 (Quasi-ML) Let $l\left(z_i, \theta\right)$ *denote the quasi*[2]*-loglikelihood function for* θ *, and let* $s\left(z_i, \theta\right) = \partial l\left(z_i, \theta\right)/\partial\theta$ *denote the quasi-score vector. Then*

$$E\left[s\left(z_i, \theta_0\right)\right] = 0.$$

[2] In the sense of White (1982)

Let

$$g(z_i,\theta) = g_i(\theta), \quad \sum_{i=1}^{n} g(z_i,\theta)/n = \widehat{g}(\theta),$$
$$\partial g(z_i,\theta)/\partial\theta = G_i(\theta), \quad \sum_{i=1}^{n} (\partial g(z_i,\theta)/\partial\theta)/n = \widehat{G}_i.$$

Typically θ_0 is estimated using the so-called analogy principle in which the expectation is replaced by its sample analogue, and the estimator is defined as

$$\left\|\widehat{g}\left(\widehat{\theta}\right)\right\| = 0$$

We shall call this estimator a Z estimator. Consistency of $\widehat{\theta}$ and asymptotic normality of $n^{1/2}\left(\widehat{\theta}-\theta_0\right)$ can be established under mild regularity conditions.

Theorem 1 *Assume that (I) the parameter space* Θ *is a compact set, (II) for all* $\zeta > 0$ $\inf_{\|\theta-\theta_0\|>\zeta} \|Eg(\theta)\| \geq \varepsilon(\zeta) > 0$, *(III)* $\sup_{\theta\in\Theta} \|\widehat{g}(\theta) - Eg(\theta)\| = o_p(1)$. *Then* $\widehat{\theta} \xrightarrow{p} \theta_0$.*Assume further that (IV)* $\theta_0 \in int\{\Theta\}$, *(V)* $\sup_{\theta\in\mathcal{N}_0} \left\|\widehat{G}(\theta) - E[G(\theta)]\right\| = o_p(1)$, *(VI)* $E(G_0)$ *is nonsingular where* $G_0 = G(\theta_0)$, *(VII)* $n^{1/2}\widehat{g}(\theta_0) \xrightarrow{d} N(0,\Omega_0)$ *where* $\Omega_0 = E\left[g(\theta_0)\, g(\theta_0)'\right]$. *Then*

$$n^{1/2}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, [E(G_0)]^{-1}\Omega_0\left[E(G_0)'\right]^{-1}\right)$$

Proof. Note that

$$\begin{aligned}\left\|Eg\left(\widehat{\theta}\right)\right\| &\leq \sup_{\theta\in\Theta}\left\|\widehat{g}\left(\theta\right)-Eg\left(\theta\right)\right\|+\left\|\widehat{g}\left(\widehat{\theta}\right)\right\| \\ &\leq o_p\left(1\right)+\left\|Eg\left(\theta_0\right)\right\|=o_p\left(1\right)\end{aligned}$$

It then follows that $\widehat{\theta}\in\left\|\theta-\theta_0\right\|<\zeta$ w.p.a. 1 and since ζ is arbitrary $\widehat{\theta}\xrightarrow{p}\theta_0$.The asymptotic normality follows by standard mean value expansion of $0=\widehat{g}\left(\widehat{\theta}\right)$.

■

Inference about θ can be based on a number of well-known statistics including Wald, Lagrange multiplier and likelihood ratio (if we are willing to specify a density for z_i).

6 EL for exactly identified moment conditions models

Here we consider how EL method can be used to test the simple hypothesis $H_0 : \theta = \theta_0$.
In the case of moment condition models the multinomial is constrained so as to satisfy (1), that is

$$\max_{\pi_i} \prod_{i=1}^{n} \pi_i \text{ s.t. } \sum_{i=1}^{n} \pi_i = 1 \text{ and } \sum_{i=1}^{n} \pi_i g_i(\theta_0) = 0. \tag{2}$$

The solution to (2) can be found by the following Lagrange multiplier argument. Let

$$\mathcal{L}(\pi_i, \lambda, \gamma) = \sum_{i=1}^{n} \log \pi_i - \lambda' \sum_{i=1}^{n} \pi_i g_i(\theta_0) + \gamma \left(\sum_{i=1}^{n} \pi_i - 1 \right)$$

denote the Lagrangian, and let

$$\begin{aligned}
\partial \mathcal{L}(\pi_i, \lambda, \gamma) / \partial p_i &= 1/\pi_i - \widehat{\lambda}' g_i(\theta_0) + \widehat{\gamma} = 0, \\
\partial \mathcal{L}(\pi_i, \lambda, \gamma) / \partial \lambda &= \sum_{i=1}^{n} \pi_i g_i(\theta_0) = 0, \\
\partial \mathcal{L}(\pi_i, \lambda, \gamma) / \partial \gamma &= \sum_{i=1}^{n} \pi_i - 1 = 0.
\end{aligned}$$

Summing over the i the first line and multiplying by π_i yields

$$\begin{aligned}\sum_{i=1}^{n} 1/\pi_i - \widehat{\lambda}' \sum_{i=1}^{n} g_i(\theta_0) + \sum_{i=1}^{n} \widehat{\gamma} &= 0 \Rightarrow \sum_{i=1}^{n} (1 + \widehat{\gamma}\pi_i) = 0 \\ \widehat{\gamma} &= -n.\end{aligned}$$

Replacing this to the first line and multiplying both sides by π_i gives

$$\widehat{\pi}_i^{-1} = n\left(1 + \widehat{\lambda}' g_i(\theta_0)\right). \tag{3}$$

The EL ratio for the null hypothesis $H_0 : \theta = \theta_0$ is then

$$R(\theta_0) = \prod_{i=1}^{n} n\widehat{\pi}_i = \prod_{i=1}^{n} \left(1 + \widehat{\lambda}' g_i(\theta_0)\right)^{-1}.$$

Let

$$W(\theta_0) = -2\log R(\theta_0) = 2\sum_{i=1}^{n} \log\left(1 + \widehat{\lambda}' g_i(\theta_0)\right)$$

denote the (log) EL ratio test statistic. The following theorem is the nonparametric likelihood version of Wilks's (1938) theorem adapted from Owen (1988) and Owen (1990) to moment conditions models.

Theorem 2 Assume that (I) $0 \in ch\left\{g_1\left(\theta_0\right), ..., g_n\left(\theta_0\right)\right\}$, *(II)*$E\left\|g\left(\theta_0\right)\right\|^2 < \infty$, *(III)*$E\left[g\left(\theta_0\right) g\left(\theta_0\right)'\right] = \Omega_0$ *p.d.. Then under* H_0

$$W\left(\theta_0\right) \xrightarrow{d} \chi^2_k.$$

Proof. First we establish the convergence rate for the Lagrange multiplier $\widehat{\lambda}$.
By definition $\tilde{\lambda}$ solves

$$\sum_{i=1}^{n} \widehat{\pi}_i g_i\left(\theta_0\right) = \sum_{i=1}^{n} g_i\left(\theta_0\right)/n \left(1 + \widehat{\lambda}' g_i\left(\theta_0\right)\right) = 0. \tag{4}$$

Let $\widehat{\lambda} = \eta\rho$ where $\|\eta\| = 1$ and $\rho > 0$. Then (4) can be written as

$$0 = \widehat{g}\left(\theta_0\right) - \sum_{i=1}^{n} g_i\left(\theta_0\right) g_i\left(\theta_0\right)' \eta\rho/n \left(1 + \eta'\rho g_i\left(\theta_0\right)\right).$$

Multiplying the right hand side by η' and rearranging

$$0 = \eta' \sum_{i=1}^{n} g_i\left(\theta_0\right)\left(1 + \eta'\rho g_i\left(\theta_0\right)\right)/n - \eta'\widehat{\Omega}\left(\theta_0\right)\eta\rho/n$$

and note that $\sigma_{\min}\left(\Omega_0\right) \leq \eta'\widehat{\Omega}\left(\theta_0\right)\eta \leq \sigma_{\max}\left(\Omega_0\right)$ w.p.a.1, and that $\|\widehat{g}\left(\theta_0\right)\| = O_p\left(n^{-1/2}\right)$ by CLT and

$\max_i \|g_i(\theta_0)\| = o_p\left(n^{1/2}\right)$ by Borel-Cantelli lemma so that

$$\begin{aligned} 0 &\leq \|\eta\| \|\widehat{g}(\theta_0)\| \left(1 + \|\eta\| \rho \max_i \|g_i(\theta_0)\|\right) - \sigma_{\min}(\Omega_0)\rho \\ 0 &\leq O_p\left(n^{-1/2}\right)\left(1 + \rho o_{p.}\left(n^{1/2}\right)\right) - \sigma_{\min}(\Omega_0)\rho \end{aligned}$$

which shows that $\rho = O_p\left(n^{-1/2}\right)$ i.e. $\widehat{\lambda} = O_p\left(n^{-1/2}\right)$.
Next we find a stochastic approximation for $\widehat{\lambda}$.
Using again (4) and noting that $\max_i \left|\widehat{\lambda}' g_i(\theta_0)\right| = o_p(1)$ we have

$$\begin{aligned} 0 &= \widehat{g}(\theta_0) - \sum_{i=1}^{n} g_i(\theta_0) g_i(\theta_0)' \widehat{\lambda}/n \left(1 + \widehat{\lambda}' g_i(\theta_0)\right) \Rightarrow \\ \widehat{\lambda} &= \widehat{\Omega}(\theta_0)^{-1} \widehat{g}(\theta_0) + o_p(1) = \Omega_0^{-1} \widehat{g}(\theta_0) + o_p(1). \end{aligned} \tag{5}$$

Finally by Taylor expansion

$$\begin{aligned} W(\theta_0) &= 2\sum_{i=1}^{n} \log\left(1 + \widehat{\lambda}' g_i(\theta_0)\right) = 2\widehat{\lambda}' \sum_{i=1}^{n} g_i(\theta_0) - \\ &\quad \widehat{\lambda}' \sum_{i=1}^{n} g_i(\theta_0) g_i(\theta_0)' \widehat{\lambda} + O_p\left(\sum_{i=1}^{n} \left(\widehat{\lambda}' g_i(\theta_0)\right)^3\right). \end{aligned} \tag{6}$$

Replacing (5) in (6) and noting that

$$\sum_{i=1}^{n}\left(\widehat{\lambda}'g_i\left(\theta_0\right)\right)^3 \leq \max_i\left|\widehat{\lambda}'g_i\left(\theta_0\right)\right|\widehat{\lambda}\sum_{i=1}^{n}g_i\left(\theta_0\right)g_i\left(\theta_0\right)'\widehat{\lambda} = o_p\left(1\right)$$

we get

$$W\left(\theta_0\right) = n\widehat{g}\left(\theta_0\right)'\Omega_0^{-1}\widehat{g}\left(\theta_0\right) + o_p\left(1\right)$$

from which the conclusion follows by CMT. ■

Remark 1 The convex hull condition (I) has an intuitive interpretation if the observations are assumed to be univariate. In this case it means that 0 *is assumed to be contained in the interval between the minimum and the maximum of the observations.*

Remark 2 In practice Theorem 4 can be used as follows: suppose that we have a simple hypothesis of interest in an exactly identified moment condition model. Then assume that λ *is a free-varying parameter and note that the null hypothesis* $H_0 : \theta = \theta_0$ *can be reformulated in terms of* $H_0 : \lambda = 0$*. The latter is the dual of the original hypothesis, and* λ *is the Lagrange multiplier associated with the sample moment condition* $\sum_{i=1}^{n}\pi_i g_i\left(\theta_0\right) = 0$*. The resulting test statistic is given by twice the maximised function* $\sum_{i=1}^{n}\log\left(1+\lambda'g_i\left(\theta_0\right)\right)$ *with respect to* λ*.*
This is the idea behind a number of artificial likelihoods including

1. Mykland's (1995) dual likelihood
2. Smith's (1997) generalised empirical likelihood
3. Chesher & Smith's (1997) augmented densities for specification testing.

7 EL for overidentified moment condition models

Assume that $\dim(g(\cdot)) = l \geq \dim(\theta) = k$. As before

$$E[g_i(\theta_0)] = 0 \tag{7}$$

for a unique θ_0.

Example 3 (Generalised instrumental variable regression). Let $z_i = [y_i, x_i', w_i']'$, and let $y_i = x_i'\theta_0 + \varepsilon_i$ where ε_i is an unobservable error. Suppose that the errors are not orthogonal to the regressors, but there exists an l-dimensional vector of instrumental variables w_i satisfying $rank(E(w_i x_i')) = k$ and $E(w_i\varepsilon_i) = 0$. The latter property defines a moment condition model with

$$E(w_i\varepsilon_i) = E[w_i(y_i - x_i'\theta_0)] = 0.$$

Typically θ_0 is estimated using Hansen's (1982) generalised method of moment (GMM) method, and the GMM (or generalised Z) estimator is defined as

$$\left\|\widehat{g}\left(\widehat{\theta}_{GMM}\right)\right\|_{\widehat{W}} = \inf_{\theta\in\Theta} \|\widehat{g}(\theta)\|_{\widehat{W}}$$

where $\widehat{W}$ is a possibly random positive semidefinite $l \times l$ matrix.
Under essentially the same assumptions as those of Theorem 1 it is possible to show that

$$n^{1/2}\left(\widehat{\theta}_{GMM} - \theta_0\right) \xrightarrow{d} N\left(0, \left(G_0' W G_0\right)^{-1} G_0' W \Omega_0 W G_0 \left(G_0' W G_0\right)^{-1}\right) \tag{8}$$

where $W = p\lim\left(\widehat{W}\right)$.

Remark 3 Note that (8) *crucially depends on* W*. Hansen (1982) (see also Chamberlain (1987)) showed that the optimal (in the sense of smallest possible variance) choice of* W *is* Ω_0^{-1}*. In this case we obtain the so-called efficient GMM estimator, that is*

$$n^{1/2}\left(\widehat{\theta}_{GMM} - \theta_0\right) \xrightarrow{d} N\left(0, \left(G_0' \Omega_0^{-1} G_0\right)^{-1}\right). \tag{9}$$

Hansen (1982) also devised a general misspecification test based on the so-called J statistic

$$J_n\left(\widehat{\theta}_{GMM}\right) = n\widehat{g}\left(\widehat{\theta}_{GMM}\right)' \widehat{\Omega}\left(\widehat{\theta}_{GMM}\right)^{-1} \widehat{g}\left(\widehat{\theta}_{GMM}\right) \xrightarrow{d} \chi^2_{l-k}. \tag{10}$$

Suppose now that we are interested to test the hypothesis

$$H_0 : h(\theta_0) = 0$$

where $h(\cdot)$ is an $\mathbb{R}^p$ valued vector of continuously differentiable functions, and assume that $rank(H(\theta)) = p$ where $H(\theta) = \partial H(\theta)/\partial\theta'$. Let

$$\left\| \widehat{g}\left(\widehat{\theta}^c_{GMM}\right) \right\|_{\widehat{W}} = \inf_{\theta\in\Theta} \|\widehat{g}(\theta)\|_{\widehat{W}} \quad \text{s.t. } h(\theta) = 0$$

denote the constrained GMM estimators, and let

$$\begin{aligned} D &= J_n\left(\widehat{\theta}_{GMM}\right) - J_n\left(\widehat{\theta}^c_{GMM}\right) \\ LM &= n(\widehat{\gamma}^c)'\widehat{\Phi}\left(\widehat{\theta}^c_{GMM}\right)\widehat{\gamma}^c,\ W = nh\left(\widehat{\theta}_{GMM}\right)'\widehat{\Phi}\left(\widehat{\theta}_{GMM}\right)^{-1}h\left(\widehat{\theta}_{GMM}\right), \\ \widehat{\Phi}(\cdot) &= H(\cdot)\left(\widehat{G}(\cdot)'\widehat{\Omega}(\cdot)^{-1}\widehat{G}(\cdot)\right)^{-1}H(\cdot)' \end{aligned} \tag{11}$$

denote the GMM based distance (D), Lagrange multiplier (LM) and Wald (W) test statistics for H_0, where $\widehat{\gamma}^c$ is a vector of Lagrange multipliers. By standard arguments (see for example Newey & McFadden (1994)) it is not difficult to see that

$$D, LM, W \xrightarrow{p} \chi^2_p.$$

We now show that EL can be used in overidentified moment condition models to obtain estimators and

test statistics that are asymptotically equivalent to those based on the efficient GMM method.
Note that because of the overidentification we have also to estimate the unknown θ_0.
Thus the estimation process becomes more complicated as the EL estimator $\widehat{\theta}$ is defined as

$$\widehat{\theta} = \max_{\theta \in \Theta} \sup_{\lambda} - \sum_{i=1}^{n} \log\left(1 + \lambda' g_i\left(\theta\right)\right)$$

(This is what Qin & Lawless (1994) call maximum empirical likelihood estimator -MELE).
The following theorem (an i.i.d. version of Theorem 1 of Kitamura (1997)) shows that the EL estimator has the same asymptotic distribution as that of the efficient GMM estimator. Let $\Gamma\left(\theta, \delta\right)$ denote an open sphere centred at θ with radius δ.

Theorem 3 Assume that (I) the parameter space Θ *is compact, (II) for a small* $\delta > 0$ $E \sup_{\theta^* \in \Gamma(\theta^*, \delta)} - \log\left(1 + \lambda' g_i\left(\theta^*\right)\right) < \infty$*, (III)* $\theta_0 \in int\left(\Theta\right)$*, (IV)* $g\left(z_i, \theta\right)$ *is twice continuously differentiable at* θ_0*, (V)* Ω_0 *is p.d. (VI)* $E \sup_{\theta^* \in \Gamma(\theta^*, \delta)} \left\| g\left(\theta^*\right)\right\|^{2+\epsilon} < \infty$ *for some* $\epsilon > 0$*,* $E \sup_{\theta^* \in \Gamma(\theta^*, \delta)} \left\| G\left(\theta^*\right)\right\|^2 < \infty$*,* $E \sup_{\theta^* \in \Gamma(\theta^*, \delta)} \left\| \partial G_i\left(\theta^*\right) / \partial \theta_j \right\| < \infty \ (j = 1, ..., k)$*. Then* $\widehat{\lambda} \xrightarrow{p} 0$*,* $\widehat{\theta} \xrightarrow{p} \theta_0$*, and*

$$n^{1/2} \begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta} - \theta_0 \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Psi_0 & 0 \\ 0 & \left(G_0' \Omega_0^{-1} G_0\right)^{-1} \end{bmatrix} \right), \tag{12}$$

where $\Psi_0 = \Omega_0^{-1}\left(I - G_0\left(G_0' \Omega_0^{-1} G_0\right)^{-1} G_0' \Omega_0^{-1}\right)$

Proof. The consistency of $\widehat{\lambda}$ follows as in the proof of Theorem 2 which shows that $\widehat{\lambda} = O_p\left(n^{-1/2}\right)$. The proof of consistency of $\widehat{\theta}$ is based on the classical argument used by Wald. In particular the proof consists of two steps: first to check that outside of an arbitrary neighbourhood containing θ_0 the sample objective function is bounded away from the maximum the population objective function achieved at θ_0. Second that the maximum of the sample objective function is not smaller than its value at θ_0. Since the latter converges to its expectation it follows that the maximum belongs to the arbitrary neighbourhood containing θ_0, which proves consistency. To be specific note that since the "optimal" λ is actually 0 we have

$$E - \log\left(1 + \lambda' g_i(\theta)\right) \leq 0 \tag{13}$$

for all $\theta \neq \theta_0$. Moreover by (II)

$$\lim_{\delta \to 0} E \sup_{\theta^* \in \Gamma(\theta, \delta)} -\log\left(1 + \lambda' g_i(\theta^*)\right) = E - \log\left(1 + \lambda' g_i(\theta)\right). \tag{14}$$

Because the parameter space Θ is compact we can cover the set $\Theta(\delta) = \Theta - \Gamma(\theta_0, \delta)$ with $\Gamma(\theta_j, \delta)$ $(j = 1, ..., h)$ so that by (14) and $H_j > 0$

$$E \sup_{\theta^* \in \Gamma(\theta_j, \delta)} -\log\left(1 + \lambda' g_i(\theta^*)\right) = -2H_j$$

whence by LLN for all j

$$\Pr\left(\sum_{i=1}^{n}\sup_{\theta^{*}\in\Theta(\delta)}-\log\left(1+\lambda' g_i\left(\theta^{*}\right)\right)/n>-H\right)<\varepsilon/2$$

for $H=\min_j H_j$. At the same time note that by (13)

$$\Pr\left(\sum_{i=1}^{n}-\log\left(1+\lambda' g_i\left(\theta_0\right)\right)/n<-H\right)<\varepsilon/2$$

which shows that $\widehat{\theta}\notin\Theta\left(\delta\right)$ $w.p.a.1$ or $\Pr\left(\widehat{\theta}\in\Gamma\left(\theta_0,\delta\right)\right)\geq 1-\varepsilon$, so that the consistency follows since δ is arbitrary.

The asymptotic normality follows by mean value expansion of the first order conditions for $\widehat{\lambda}$ and $\widehat{\theta}$. To be specific note that by consistency

$$\left[\begin{array}{c}0\\0\end{array}\right]=\left[\begin{array}{c}\partial W\left(\widehat{\lambda},\widehat{\theta}\right)/\partial\lambda\\ \partial W\left(\widehat{\lambda},\widehat{\theta}\right)/\partial\theta\end{array}\right]\quad w.p.a.1$$

where

$$W\left(\lambda,\theta\right)=\sum_{i=1}^{n}\log\left(1+\lambda' g_i\left(\theta\right)\right).$$

Then by ULLN

$$\begin{aligned}\begin{bmatrix} 0 \\ 0 \end{bmatrix} &= n^{1/2} \begin{bmatrix} \partial W(0,\theta_0)/\partial\lambda \\ \partial W(0,\theta_0)/\partial\theta \end{bmatrix} + \\ & n^{-1} \begin{bmatrix} \partial^2 W\left(\overline{\lambda},\overline{\theta}\right)/\partial\lambda\partial\lambda' & \partial^2 W\left(\overline{\lambda},\overline{\theta}\right)/\partial\lambda\partial\theta' \\ \partial^2 W\left(\overline{\lambda},\overline{\theta}\right)/\partial\theta\partial\lambda' & \partial^2 W\left(\overline{\lambda},\overline{\theta}\right)/\partial\theta\partial\theta' \end{bmatrix} n^{1/2} \begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta}-\theta_0 \end{bmatrix} \\ = & \begin{bmatrix} n^{1/2}\widehat{g}(\theta_0) \\ 0 \end{bmatrix} + \begin{bmatrix} \Omega_0 & G_0 \\ G_0' & 0 \end{bmatrix} n^{1/2} \begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta}-\theta_0 \end{bmatrix},\end{aligned} \tag{15}$$

and the result follows by standard manipulations, CLT and CMT. ■

Remark 4 (12) *shows an important aspect of EL method, that is that the estimators* $\widehat{\lambda}$ *and* $\widehat{\theta}-\theta_0$ *are asymptotically independent. This shows that EL is using efficiently the information provided by the overidentification.*

We now show how EL can be used to obtain inferences on θ.
We first consider the EL analogue of Hansen's (1982) J general test for misspecification (10).

Theorem 4 Under the same assumptions of Theorem 3,

$$2\sum_{i=1}^{n}\log\left(1+\widehat{\lambda}'g_i\left(\widehat{\theta}\right)\right)\overset{d}{\rightarrow}\chi^2_{l-k}. \tag{16}$$

Proof. Because $\max_i\left|\widehat{\lambda}'g_i\left(\widehat{\theta}\right)\right|=o_p(1)$ we can use a Taylor expansion about 0

$$\begin{aligned}2\sum_{i=1}^{n}\log\left(1+\widehat{\lambda}'g_i\left(\widehat{\theta}\right)\right) &= 2\sum_{i=1}^{n}\left(\widehat{\lambda}'g_i\left(\widehat{\theta}\right)-\left(\widehat{\lambda}'g_i\left(\widehat{\theta}\right)\right)^2/2\right)+O\left(\sum_{i=1}^{n}\left(\widehat{\lambda}'g_i\left(\widehat{\theta}\right)\right)^3\right)\\ &= n\widehat{g}\left(\widehat{\theta}\right)'\widehat{\Omega}\left(\widehat{\theta}\right)^{-1}\widehat{g}\left(\widehat{\theta}\right)'+o_p(1)\overset{d}{\rightarrow}\chi^2_{l-k}\end{aligned}$$

where the second line follows by $n^{1/2}\widehat{\lambda}=\widehat{\Omega}\left(\widehat{\theta}\right)^{-1}n^{1/2}\widehat{g}\left(\widehat{\theta}\right)+o_p(1)$ (as in the proof of Theorem 2). ■

Note that (16) does not require the explicit estimation of the covariance matrix Ω_0^{-1}, and this is clearly convenient when such estimation is difficult.

We now consider the general case of EL inference for the same hypothesis $H_0 : h(\theta_0) = 0$. EL shares an important similarity with ordinary likelihood (as well as GMM) in that the three classical tests (W, LM and LR -D in GMM case) are available. To be specific let

$$\widehat{\theta}^c = \min_{\theta \in \Theta} \sup_{\lambda} - \sum_{i=1}^{n} \log\left(1 + \lambda' g_i(\theta)\right) \text{ s.t. } h(\theta) = 0$$

denote the constrained EL estimator. Then we can define in analogy with (11)

$$\begin{aligned} ELR &= 2\sum_{i=1}^{n} \log\left(1 + \left(\widehat{\lambda}^c\right)' g_i\left(\widehat{\theta}^c\right)\right) - 2\sum_{i=1}^{n} \log\left(1 + \widehat{\lambda}' g_i\left(\widehat{\theta}^c\right)\right) \\ LM &= n(\widehat{\gamma}^c)' \widehat{\Phi}\left(\widehat{\theta}^c\right) \widehat{\gamma}^c, \; W = nh\left(\widehat{\theta}\right)' \widehat{\Phi}\left(\widehat{\theta}\right)^{-1} h\left(\widehat{\theta}\right), \end{aligned}$$

The following theorem shows that these three classical test statistics are asymptotically chi-squared.

Theorem 5 Under the same assumptions of Theorem 3,

$$ELR, LM, W \xrightarrow{d} \chi^2_p.$$

Proof. By a mean value expansion, ULLN, CMT and the results of Theorem 2, it is easy to see that

$W \xrightarrow{d} \chi^2_p$. Similarly a mean value expansion of FOCs from the Lagrangian

$$\mathcal{L}(\theta, \gamma) = -\sum_{i=1}^{n} \log\left(1 + \lambda' g_i(\theta)\right) - \gamma' h(\theta),$$

ULLN, CMT and standard manipulations

$$\begin{bmatrix} n^{1/2}\widehat{\lambda}^c \\ n^{1/2}\left(\widehat{\theta}^c - \theta_0\right) \\ n^{1/2}\widehat{\gamma}^c \end{bmatrix} = \begin{bmatrix} A^{-1}\left(I - B\left(B'A^{-1}B\right)^{-1}BA^{-1}\right) & A^{-1}B\left(B'A^{-1}B\right)^{-1} \\ \left(B'A^{-1}B\right)^{-1}B'A^{-1} & \left(B'A^{-1}B\right)^{-1} \end{bmatrix}^{-1} \times \begin{bmatrix} n^{1/2}\widehat{g}(\theta_0) \\ 0 \end{bmatrix},$$

where

$$A = \begin{bmatrix} -\Omega_0 & G_0 \\ G_0' & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ -H_0' \end{bmatrix},$$

from which by CLT and CMT

$$\begin{bmatrix} n^{1/2}\widehat{\lambda}^c \\ n^{1/2}\left(\widehat{\theta}^c - \theta_0\right) \\ n^{1/2}\widehat{\gamma}^c \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Xi_0 & 0 \\ 0 & \left(H_0\left(G_0'\Omega_0^{-1}G_0\right)^{-1}H_0'\right)^{-1} \end{bmatrix}\right)$$

and the result follows by CMT, noting the asymptotic independence between $\widehat{\gamma}^c$, $\widehat{\theta}^c - \theta_0$ and $\widehat{\lambda}^c$. Finally the distribution of ELR can be obtained by expanding both statistics about 0 and since $\widehat{\lambda}^c = \widehat{\Omega}\left(\widehat{\theta}^c\right)^{-1}\widehat{g}\left(\widehat{\theta}^c\right)$

$$ELR = n\widehat{g}\left(\widehat{\theta}^c\right)'\widehat{\Omega}\left(\widehat{\theta}^c\right)^{-1}\widehat{g}\left(\widehat{\theta}^c\right) - n\widehat{g}\left(\widehat{\theta}\right)'\widehat{\Omega}\left(\widehat{\theta}\right)^{-1}\widehat{g}\left(\widehat{\theta}\right)$$

which shows that $ELR = D + o_p(1)$. ■

8 Higher order asymptotic theory (I)

8.1 EL and saddlepoint approximation (SP)

Let

$$\widehat{K}(\theta_0) = \log\left(\sum_{i=1}^{n} \exp\left(\widehat{\xi}(\theta_0)' g_i(\theta_0)\right)/n\right)$$

denote the empirical cumulant generating function for the moment indicator $g_i(\theta_0)$ where (the saddlepoint) $\widehat{\xi}(\theta_0)$ satisfies

$$\sum_{i=1}^{n} g_i(\theta_0) \exp\left(\widehat{\xi}(\theta_0)' g_i(\theta_0)\right) = 0.$$

Let $\widehat{\delta} = n^{1/2}\left(\widehat{\theta} - \theta_0\right)$ and $\widehat{\Gamma}\left(\widehat{\delta}\right) = \sum_{i=1}^{n}\left(\widehat{\delta}'\widehat{G}'\widehat{\Omega}^{-1} g_i\left(\widehat{\theta}\right)\right)^3$.

Proposition 1 Under the same assumptions of Theorem 2, with (II) strengthened to $E\|g(\theta_0)\|^3 < \infty$. *Then*

$$n\widehat{K}(\theta_0) = -W(\theta_0)/2 + \widehat{\Gamma}\left(\widehat{\delta}\right)/\left(6n^{1/2}\right) + O\left(n^{-1}\right). \tag{17}$$

Proof. See Monti & Ronchetti (1993). ■

Relation (17) can be used in two ways:

- First it can be used to obtain a nonparametric approximation to the density of the Z-estimator $\widehat{\theta}$ by replacing $-W(\theta_0)/2 + \widehat{\Gamma}\left(\widehat{\delta}\right) / \left(6n^{1/2}\right)$ into the empirical saddlepoint approximation at θ_0

$$\widehat{f}(\theta_0) = (n/2\pi)^{k/2} \left|\widehat{K}(\theta_0)''\right|^{-1/2} \left|\widehat{A}(\theta_0)\right| \exp\left(n\widehat{K}(\theta_0)\right)$$

  where

- Secondly it can be used to construct accurate nonparametric confidence regions for EL by replacing $W(\theta_0)$ with $-2n\widehat{K}(\theta_0) + \widehat{\Gamma}\left(\widehat{\delta}\right) / \left(3n^{1/2}\right)$.

8.2 Bartlett corrections

- Bartlett (and more generally Bartlett-type - see for example Cribari-Neto & Cordeiro (1996) for a survey of econometric applications) corrections are designed to bring the actual size of an asymptotically χ^2 distributed test statistic $S(\theta_0)$ closer to its nominal size.

$S(\theta_0)$ is said to be Bartlett-correctable if the terms of order $O\left(n^{-1}\right)$ in the asymptotic distribution of

$$S(\theta_0)^B = S(\theta_0)/E[S(\theta_0)] = \left(1 - b(\theta_0)/n + O\left(n^{-2}\right)\right) S(\theta_0)$$

vanish because $E[S(\theta_0)] = 1 + b(\theta_0)/n + O\left(n^{-2}\right)$, this result holding for all the cumulants to an order $O\left(n^{-2}\right)$ (Barndorff-Nielsen & Hall 1988).

DiCiccio, Hall & Romano (1991) showed that the EL ratio for the so-called smooth function of means model is Bartlett-correctable.

In the case of Z-estimators Bravo (2004) shows that

$$b(\theta_0) = E\left[\sum_{j,l=1}^{k} h_j(\theta_0)^2 h_l(\theta_0)^2\right]/2 - E\left[\sum_{j,l,m=1}^{k} \left(h_j(\theta_0) h_l(\theta_0) h_m(\theta_0)\right)^2\right]/3$$

where $h_j(\theta_0)$ is the jth $(j = 1, ..., k)$ component of the vector $h(\theta_0) = \Omega_0^{-1/2} g(\theta_0)$, which can be consistently estimated by say $\widehat{b}\left(\widehat{\theta}\right)$. Let

$$W^{\widehat{B}}(\theta_0) = W(\theta_0) / \left(1 + \widehat{b}\left(\widehat{\theta}\right)/(nk)\right) \tag{18}$$

denote the Bartlett corrected EL ratio. The following theorem shows that $W^{\widehat{B}}(\theta_0)$ is third-order accurate.

Theorem 6 Assume that (I) $W(\theta_0)$ *admits a valid Edgeworth expansion (in the sense of Chandra & Ghosh (1980)), (II)* $n^{1/2}\left\|\widehat{\theta}-\theta_0\right\| = O_p(1)$*. Then, for some* $c_0 \geq 0$

$$\sup_{c\in[c_0,\infty)}\left|\Pr\left(W^{\widehat{B}}(\theta_0)\leq c\right)-\int_0^c \chi_k^2(x)\,dx\right| = O\left(n^{-2}\right). \tag{19}$$

Proof. We only sketch the key steps. First one calculates the so-called signed squared root vector say $W_{1/2}(\theta)$ that is an $\mathbb{R}^k$-valued such that

$$W_{1/2}(\theta_0)' W_{1/2}(\theta_0) = W(\theta_0) + O_p\left(n^{-3/2}\right).$$

Second by evaluating the cumulants of such vector it is possible to show that

$$W_r(\theta_0) \sim N\left(\gamma(\theta_0)/n^{1/2}, I+\Gamma(\theta_0)/n\right) + O\left(n^{-3/2}\right), \tag{20}$$

where $\gamma(\theta_0)$ and $\Gamma(\theta_0)$ are, respectively, a vector and matrix such that $\gamma(\theta_0)'\gamma(\theta_0) + trace(\Gamma(\theta_0)) = b(\theta_0)$.

Third using an Edgeworth expansion it is possible to show that the density of $W_{1/2}(\theta_0)' W_{1/2}(\theta_0)$ is pro-

portional to $\chi_k^2(x)[1+b(\theta_0)x/n]+O\left(n^{-2}\right)$ so that scaling $W(\theta_0)$ by a factor $1+b(\theta_0)/(nk)$ eliminates the coefficient of $1/n$ in the expansion of $W^B(\theta_0)$. Finally by a standard mean value expansion of $\widehat{b}\left(\widehat{\theta}\right)$ it follows that $\Pr\left(W^{\widehat{B}}(\theta_0)\leq c\right)=\Pr\left(W^B(\theta_0)\leq c\right)+O\left(n^{-3/2}\right)$ where the last term is actually of order $O\left(n^{-2}\right)$ by the symmetry argument of Barndorff-Nielsen & Hall (1988). ■

Remark 5 The above theorem is valid without nuisance parameters. Recently Chen & Cui (2006) and Chen & Cui (2007) showed that it is possible to obtain Bartlett corrected statistics with nuisance parameters and for GMM estimators in overidentified moment conditions. Both results rely on a transformation of the moment indicator (and on extraordinarily heavy amount of algebra). Here we briefly consider the case of GMM estimators. Let

$$w_i(\theta)=\Psi\Omega(\theta)^{-1/2}g_i(\theta)$$

where Ψ is an $l\times l$ orthogonal matrix such that $\Psi\Omega(\theta)^{-1/2}G(\theta)U=[\Lambda,0]$ where U is a $k\times k$ orthogonal matrix, Λ is a $k\times k$ diagonal matrix. Then using this reparameterisation it is possible to obtain a third-order stochastic expansion for the ELR test statistic for the simple hypothesis $H_0:\theta=\theta_0$

$$W(\theta_0)=2\sum_{i=1}^{n}\log\left(1+\left(\widehat{\lambda}^c\right)'g_i(\theta_0)\right)-2\sum_{i=1}^{n}\log\left(1+\widehat{\lambda}'g_i\left(\widehat{\theta}\right)\right)$$

that depends on the derivatives of $\log(\cdot)$ and of $g_i(\cdot)$. Then using the same approach as that of

Theorem 6 Chen & Cui (2007) show that the squared root $W_{1/2}(\theta_0)$ has the same cumulant behaviour as that of the exactly identified model. Thus it is possible to show that

$$\sup_{c\in[c_0,\infty)}\left|\Pr\left(W^{\widehat{B}}(\theta_0)\leq c\right)-\int_0^c \chi_k^2(x)\,dx\right|=O\left(n^{-2}\right)$$

where $W^{\widehat{B}}(\theta_0)$ is as in (18) but the scalar $b(\theta_0)$ is a very complicated (and long -13 lines long!) expression. In fact they suggest to use the bootstrap to find a consistent estimator for it.

Barndorff-Nielsen, O. E. & Hall, P. (1988), 'On the level-error after Bartlett adjustment of the likelihood ratio statistic', *Biometrika* **75**, 374–378.

Bravo, F. (2003), 'Second-order power comparisons for a class of nonparametric likelihood-based tests', *Biometrika* **90**, 881–890.

Bravo, F. (2004), 'Empirical likelihood based inference with applications to some econometric models', *Econometric Theory* **20**, 231–264.

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

Chandra, T. & Ghosh, J. (1980), 'Valid asymptotic expansions for the likelihood ratio and other statistics under contiguous alternatives', *Sankhyā A* **42**, 170–184.

Chen, S. & Cui, H. (2006), 'On Bartlett correction of empirical likelihood in presence of nuisance parameters', *Biometrika* **93**, 215–220.

Chen, S. & Cui, H. (2007), On second order properties of empirical likelihood with moment restrictions. Forthcoming Journal of Econometrics.

Chesher, A. & Smith, R. J. (1997), 'Likelihood ratio specification tests', *Econometrica* **65**, 627–646.

Cribari-Neto, F. & Cordeiro, G. M. (1996), 'On Bartlett and Bartlett-type corrections', *Econometric Reviews* **15**, 339–367.

DiCiccio, T., Hall, P. & Romano, J. (1991), 'Empirical likelihood is Bartlett-correctable', *Annals of Statistics,* **19**, 1053–1061.

Hansen, L. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**, 1029–1054.

Kaplan, E. & Meier, P. (1958), 'Nonparametric estimation from incomplete data', *Journal of the American Statistical Association* **53**, 457–481.

Kitamura, Y. (1997), 'Empirical likelihood methods with weakly dependent processes', *Annals of Statistics* **25**, 2084–2102.

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

Kitamura, Y. (2006), Empirical likelihood methods in econometrics: Theory and practice. Cowles Foundations for Research in Economics, Discussion Paper 1569.

Monti, A. & Ronchetti, E. (1993), 'On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators', *Biometrika* **80**, 329–338.

Mykland, P. (1995), 'Dual likelihood', *Annals of Statistics,* **23**, 396–421.

Newey, W. & McFadden, D. (1994), Large sample estimation hypothesis testing, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics, Vol. IV', Amsterdam: North Holland.

Owen, A. (1988), 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika* **36**, 237–249.

Owen, A. (1990), 'Empirical likelihood ratio confidence regions', *Annals of Statistics* **18**, 90–120.

Owen, A. (2001), *Empirical Likelihood*, Chapman and Hall.

Qin, J. & Lawless, J. (1994), 'Empirical likelihood and general estimating equations', *Annals of Statistics* **22**, 300–325.

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

Thomas, D. & Grunkemeier, G. (1975), 'Confidence interval estimation of survival probabilities of censored data', *Journal of the American Statistical Society* **70**, 865–871.

Vardi, Y. (1982), 'Nonparametric estimation in presence of length bias', *Annals of Statistics* **10**, 616–620.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50**, 1–25.

Wilks, S. (1938), 'The large-sample distribution of the likelihood ratio for testing composite hypotheses', *Annals of Mathematical Statistics* **9**, 60–62.