Empirical likelihood methods with application to econometrics

Francesco Bravo
Lecture 2

1 Empirical discrepancies

EL estimator can be thought of as the minimiser of the “likelihood” distance between the empirical distribution and the distribution supported on the sample satisfying a given constraint.

Other distances or discrepancies could be considered. For example

\[ \sum_{i=1}^{n} \pi_i \log \left( \frac{\pi_i}{n} \right), \quad \sum_{i=1}^{n} \left( n\pi_i - 1 \right)^2 \]

correspond, respectively, to the Kullback-Liebler and Euclidean distance.

Baggerly (1998) introduced a general class of nonparametric likelihoods by considering the so-called Cressie-Read (CR henceforth) divergence between \( \pi_i \) and \( 1/n \), that is

\[ \sum_{i=1}^{n} \frac{\left( \frac{\pi_i}{n} \right)^{-\gamma} - 1}{\gamma (\gamma + 1)} \]

where \( \gamma \in \mathbb{R} \) is a user-specific parameter. In particular for \( \gamma = -2 \) one obtains the
Euclidean likelihood, for $\gamma = -1$ the Kullback-Liebler, and for $\gamma = 0$ the empirical likelihood ratio statistics$^1$.

Baggerly (1998) shows that the CR can be used to obtain asymptotically valid inference for a mean. His result can be readily generalised to the case of identified moment conditions models. Using a similar Lagrange multiplier argument as that used in Lecture 1 it is possible to show that the CR test statistic for $H_0 : \theta = \theta_0$ is

$$CR(\theta_0) = \frac{2}{\gamma(\gamma + 1)} \sum_{i=1}^{n} \left[ \left( 1 + \tilde{\zeta} + \tilde{\lambda} g_i(\theta_0) \right)^{\gamma/(\gamma+1)} - 1 \right],$$

(1)

where Lagrange multipliers $\tilde{\zeta} \in \mathbb{R}$ and $\tilde{\lambda} \in \mathbb{R}^k$ are determined by the constraints

$$\sum_{i=1}^{n} \pi_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} \pi_i g_i(\theta_0) = 0$$

Then Baggerly (1998)

$$CR(\theta_0) \xrightarrow{d} \chi_k^2.$$

$^1$ Note that the two degenerate cases $\gamma = -1$ and $\gamma = 0$ are handled by taking the limits.
All members of the CR family enjoy the following desirable statistical properties:
1. they yield convex confidence regions (at least for a multivariate mean) whose shape is typically data-determined.
2. CR regions are transformation invariant, do not require estimation of scale.

Remark 1 Interestingly the choice of $\gamma$ determines the shape of CR confidence regions and this might have some practical consequences for the resulting inference. To be specific, for $\gamma \geq 0$ the confidence regions are constrained within the convex hull of the data, and this could become a limitation when the dimension of $\mu$ is large and the sample size is small. Furthermore, the closer $0$ is to the convex hull of the data the larger the CR statistic becomes for all $\gamma \geq 0$. In the limit, if $0$ is on the convex hull of the data, the resulting confidence regions for $\gamma \geq 0$ are unbounded since the CR statistic diverges to infinity as some of the $\hat{\pi}_i$ are zero. Also for $\gamma < 0$ the CR statistic is always finite (i.e. confidence regions are always bounded) regardless of how close $0$ is
to the convex hull of the data, whereas for $\gamma < -1$ the confidence regions are allowed to extend beyond the convex hull, since negative values of $\hat{\pi}_i$ are allowed.

Imbens & Spady (2002) used the CR family to construct alternative estimators and associated statistics to those based on GMM. Corcoran (1998) proposed a very large class of statistics based on minimising the “discrepancy” function $d(x, y) (x, y \in \mathbb{R})$ that is a function such that

\[
\begin{align*}
h(x, x) &= 0, \\
\partial^r h(x, y) / \partial x^r |_{x=y} &= O_p(n^r / k_h) \text{ for } r = 1, \ldots, 4 \\
\partial^2 h(x, y) / \partial x^2 |_{x=y} &\neq 0
\end{align*}
\]

and $k_h$ is a normalising constant which depends on $h$ and is chosen so that the resulting test statistic is $O_p(1)$ as $n \to \infty$.

The empirical discrepancy approach for testing the validity of the moment condition (i.e. $H_0 : \theta = \theta_0$) is based on

\[
\min_{\pi_i} \sum_{i=1}^{n} k_h d(\pi_i, 1/n) \text{ s.t. } \sum_{i=1}^{n} \pi_i = 1 \text{ and } \sum_{i=1}^{n} \pi_i g(\theta_0) = 0.
\]
2 Generalised empirical likelihood (GEL)

An alternative very general class of statistics is based on the generalised empirical likelihood (GEL) approach of Smith (1997) and Newey & Smith (2004). To describe GEL let $\rho(v)$ be a concave real valued function with domain $V$, an open interval containing 0. Let $\Lambda_n(\theta) = \{ \lambda : \lambda' g_i(\theta) \in V, i = 1, \ldots, n \}$. The GEL estimator is defined as

$$\hat{\theta} = \min_{\theta \in \Theta} \sup_{\lambda} \hat{P}_\rho(\theta, \lambda),$$

where

$$\hat{P}_\rho(\theta, \lambda) = \sum_{i=1}^n \rho(\lambda' g_i(\theta)) / n, \quad (2)$$

that is the GEL estimator is the solution of a saddlepoint problem.

**Example 1** For $\rho(v) = \log (1 - v)$ and $V = (-\infty, 1]$ one gets EL estimator, for $\rho(v) = -\exp(v)$ and $V = (-\infty, \infty)$ one gets Efron (1981) exponential tilting estimator (see Kitamura & Stutzer (1997) and Imbens, Spady & Johnson (1998)), for $\rho(v) = -(1 + v)^2 / 2$
and \( V = (-\infty, \infty) \) one gets the continuous updating GMM estimator of Hansen, Heaton & Yaron (1996), for \( \rho(v) = -(1 + v)^{(1+\delta)/\delta} / (1 + \delta) \) where \( \delta \in \mathbb{R} \) and \( V \) depends on \( \delta \) one gets (after a reparameterisation) the CR statistic defined in (1).

The GEL criterion function (2) may be interpreted as an adaptation of the approach taken in Chesher & Smith (1997) to moment conditions models. Chesher & Smith (1997) cast tests for moment conditions into a fully parametric framework by augmenting the density of the model under the null hypothesis multiplicatively using a “carrier” function \( c(\cdot) \) which incorporates the moment conditions. For moment conditions models there is no knowledge of the distribution of the data, but notice that we can use the empirical distribution function instead. Following this “quasi-likelihood” interpretation it follows that \( \hat{P}_\rho(\theta, 0) \) corresponds to the imposition of the restriction that the auxiliary parameter \( \lambda = 0 \), which is the dual of \( H_0 : \theta = \theta_0 \).
Remark 2 Despite being asymptotically equivalent there is an important conceptual difference between the CR (and more generally the MD) and the GEL approach to estimation and inference in moment conditions models. The latter is based on introducing an auxiliary vector of free varying parameters (associated with the parameters of interest) and a criterion function over which both set of parameters are maximised. The former is based on the idea that in a fully nonparametric context the EDF does not exploit the auxiliary information available in the form of moment conditions. There is however an interesting connection relating CR and GEL in terms of duality: for each GEL estimator there is a CR estimator. This duality is useful because it shows how the computationally less complex GEL can be related to a CR estimator. Also duality justifies the interpretation as MD of the so-called implied probabilities $\hat{\pi}_i$ to be discussed later on.
Theorem 1 (Newey & Smith (2004)) Assume that (I) the parameter space $\Theta$ is a compact set, (II) $E \left[ \sup_{\theta \in \Theta} \| g(\theta) \|^\alpha \right] < \infty$ for some $\alpha > 2$, (III) $\rho(v)$ is twice continuously differentiable in a neighbourhood of 0, (IV) $\Omega_0$ is p.d., (V) $\theta_0 \in \text{int} \{ \Theta \}$, (VI) $E \left[ \sup_{\theta \in \mathcal{N}_0} \| G(\theta) \| \right] < \infty$ (VII) (a) $\text{rank} (G_0) = k$ and (b) $G_0' \Omega_0^{-1} G_0$ is nonsingular. Then

$$n^{1/2} \left[ \begin{array}{c} \hat{\lambda} \\ (\hat{\theta} - \theta_0) \end{array} \right] \xrightarrow{d} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Psi_0 & 0 \\ 0 & (G_0' \Omega_0^{-1} G_0)^{-1} \end{bmatrix} \right),$$

$$2n \left( \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) - \hat{P}_\rho (0) \right) \xrightarrow{d} \chi^2_{l-k}.$$

Proof. First we show the consistency of both $\hat{\lambda}$ and $\hat{\theta}$.

Let $\Lambda_n = \{ \lambda : \| \lambda \| \leq n^{-\beta} \}$ for $1/\alpha \leq \beta < 1/2$. By the Borel-Cantelli lemma $\max_i \| g_i (\theta_i) \| = o_{a.s.} \left( n^{1/\alpha} \right)$ and thus

$$\sup_{\theta \in \Theta, \lambda \in \Lambda_n} \max_i \| \lambda' g_i (\theta_i) \| = o_p(1) \text{ and } \Lambda_n \subseteq \hat{\Lambda}_n (\theta) \text{ w.p.a.1.}$$
Then by (III) and a Taylor expansion

\[
\sup_{\lambda \in \Lambda_n(\theta_0)} \tilde{P}_\rho(\theta_0, \lambda) \leq \rho_0 - \tilde{\lambda}^t \tilde{g}(\theta) + \tilde{\lambda}^t \sum_{i=1}^n \rho_2 \left( \tilde{\lambda}^t g_i(\theta_0) \right) g_i(\theta_0) g_i(\theta_0)^t \tilde{\lambda} / (2n) \\
\leq \rho_0 - \|\tilde{\lambda}\| \|\tilde{g}(\theta_0)\| - C \|\tilde{\lambda}\|^2.
\]

(3)

Assume that \(\|\tilde{g}(\theta_0)\| = O_p \left( n^{-1/2} \right) \) since \(\rho_0 \leq \sup_{\lambda} \tilde{P}_\rho(\theta_0, \lambda)\) we get

\[\|\tilde{\lambda}\| \leq \|\tilde{g}(\theta_0)\| = O_p \left( n^{-1/2} \right).\]

We next show that \(\|\tilde{g}(\hat{\theta})\| = O_p \left( n^{-1/2} \right) \) where \(\hat{\theta} = \arg\min_{\theta \in \Theta} \tilde{P}(\lambda, \theta)\) for \(\lambda \in \Lambda_n\). Let \(\overline{\lambda} = -n^{\beta} \tilde{g}(\hat{\theta}) / \|\tilde{g}(\hat{\theta})\|\) and note that by a similar expansion as that given in (3)

\[\tilde{P}_\rho(\hat{\theta}, \overline{\lambda}) \geq \rho_0 + n^{-\beta} \|\tilde{g}(\hat{\theta})\| - C n^{-2\beta}.\]

(4)

Combining (3) and (4) and using the definition of \(\hat{\theta}\) and the fact that \(\tilde{\lambda}\) is a saddlepoint...
gives
\[ \rho_0 + n^{-\beta} \left\| \hat{g} \left( \hat{\theta} \right) \right\| - C n^{-2\beta} \leq \hat{P}_\rho \left( \hat{\theta}, \lambda \right) \leq \sup_{\lambda \in \Lambda_n(\theta_0)} \hat{P}_\rho \left( \hat{\theta}, \lambda \right) \leq \rho_0 + O_p \left( n^{-1} \right), \quad (5) \]
so that rearranging \[ \left\| \hat{g} \left( \hat{\theta} \right) \right\| \leq O_p \left( n^{-\beta} \right). \] Let \( \tilde{\lambda} = -\varepsilon_n \hat{g} \left( \hat{\theta} \right) \) where \( \varepsilon_n \to 0 \), so that \( \tilde{\lambda} \in \Lambda_n \). Then (5) shows that
\[ \rho_0 - \left\| \tilde{\lambda} \right\| \left\| \hat{g} \left( \hat{\theta} \right) \right\| - C \left\| \tilde{\lambda} \right\|^2 = \rho_0 - \varepsilon_n \left\| \hat{g} \left( \hat{\theta} \right) \right\|^2 - C \varepsilon_n^2 \left\| \hat{g} \left( \hat{\theta} \right) \right\|^2 \leq \rho_0 + O_p \left( n^{-1} \right) \]
which implies that \( \left\| \hat{g} \left( \hat{\theta} \right) \right\| = O_p \left( n^{-1/2} \right) \). The consistency of \( \hat{\theta} \) follows by noting that by ULLN and T \( E \left[ g \left( \hat{\theta} \right) \right] \to 0 \) and since \( E \left[ g \left( \theta_0 \right) \right] = 0 \) \( \hat{\theta} \to^p \theta_0 \). The consistency of \( \hat{\lambda} \) follows by noting that
\[ \left\| \hat{\lambda} \right\| \leq \left\| \hat{g} \left( \hat{\theta} \right) \right\| \to^p 0. \]
The asymptotic distribution of \( n^{1/2} \left[ \hat{\lambda}', \left( \hat{\theta} - \theta_0 \right)' \right]' \) follows by mean value expansion of
the FOCs using ULLN, CMT noting that $\rho_j \left( \hat{\lambda}' g_i \left( \hat{\theta} \right) \right) \overset{p}{\rightarrow} \rho_j (0)$ for $j = 1, 2$.

The distribution of $2n \left( \hat{P}_\rho \left( \hat{\theta}, \hat{\lambda} \right) - \hat{P}_\rho (0) \right)$ can be obtained by a second order Taylor expansion about 0 noting as in the proof of Theorem 4 that

$$n^{1/2} \hat{\lambda} = \hat{\Omega} \left( \hat{\theta} \right)^{-1} n^{1/2} \hat{g} \left( \hat{\theta} \right) + o_p (1)$$

and therefore

$$2n \left( \hat{P}_\rho \left( \hat{\theta}, \hat{\lambda} \right) - \hat{P}_\rho (0) \right) = n \hat{g} \left( \hat{\theta} \right)' \hat{\Omega} \left( \hat{\theta} \right)^{-1} \hat{g} \left( \hat{\theta} \right) + o_p (1).$$
As with EL it is easy to see that GEL can be used to test the nonlinear hypothesis 
\( H_0 : h(\theta_0) = 0 \). Let

\[
\hat{\theta}^c = \min_{\theta \in \Theta} \sup_{\lambda} \hat{P}_\rho (\theta, \lambda) \text{ s.t. } h(\theta) = 0
\]
denote the constrained GEL estimator. Then we can define

\[
D_\rho = 2n \left( \hat{P}_\rho \left( \hat{\theta}, \hat{\lambda} \right) - \hat{P}_\rho \left( \hat{\theta}^c, \hat{\lambda}^c \right) \right),
\]

\[
LM_\rho = n (\hat{\gamma}^c)' \hat{\Phi} (\hat{\theta}) \hat{\gamma}^c, \quad W_\rho = nh \left( (\hat{\theta})' \hat{\Phi} (\hat{\theta})^{-1} \right) ^{-1} h \left( \hat{\theta} \right).
\]

**Theorem 2** Under the same assumptions of Theorem 1

\[
D_\rho, LM_\rho, W_\rho \xrightarrow{d} \chi_p^2.
\]

**Proof.** Almost identical to that of EL. ■
3 Higher order asymptotic theory (II)

- Compare GEL estimators and test statistics: higher order bias and efficiency, and local power.

Concerning the former the following theorem (Theorems 4.1, 4.2 and Corollary 4.3 of Newey & Smith (2004)) establishes an optimality property of the EL in terms of bias.

**Theorem 3** Assume that (I)-(VI) of Theorem 1 hold. Furthermore assume that (VII) 
\[ E \sup_{\theta \in \mathcal{N}_0} \left\| \frac{\partial^k g(\theta)}{\partial \theta^{j_1} \ldots \partial \theta^{j_k}} \right\|^6 < \infty \]  
for \( k = 1, \ldots, 4 \), (VIII) for each \( \beta \in \mathcal{N} \), 
\[ \left\| \frac{\partial^4 g(\theta)}{\partial \theta^{j_1} \ldots \partial \theta^{j_4}} - \frac{\partial^4 g(\theta_0)}{\partial \theta^{j_1} \ldots \partial \theta^{j_4}} \right\| \leq b(z) \|\theta - \theta_0\| \]
and \( E [b(z)]^6 < \infty \), (IX) \( \rho(\cdot) \) is four times continuously differentiable with Lipschitz derivative in a neighbourhood of 0, (X) \( \widehat{W} = W + \sum_{i=1}^{n} \xi(z_i)/n + O_p\left(n^{-1}\right) \) where
\( E \left[ \xi \left( z_i \right) \right] = 0, \quad E \left[ \left\| \xi \left( z_i \right) \right\|^6 \right] < \infty. \) Then

\[
\text{Bias} \left( \hat{\theta}_{GMM} \right) = \sum_{j=1}^{4} B_j
\]

\[
\text{Bias} \left( \hat{\theta}_{GEL} \right) = B_1 + (1 + \rho_3/2) B_2
\]

\[
\text{Bias} \left( \hat{\theta}_{EL} \right) = B_1
\]

where each \( B_j \) is a rather complicated function involving expectations of the first two derivatives of \( g \left( \cdot \right) \), \( \Omega \) and the limiting weight matrix \( W \) (for GMM).

**Remark 3** Note that each of the \( B \) terms has an interesting interpretation: \( B_1 \) is the bias for a GMM estimator with optimal linear combination \( G' \Omega^{-1} g \left( \theta \right) \), \( B_2 \) and \( B_3 \) arise from the estimation of \( G \) and \( \Omega \), respectively and \( B_4 \) arises from the choice of preliminary estimator for the weight matrix. The bias of GEL estimators does not depend on...
the estimation of $G$ nor on that of the preliminary estimator. Note that for EL $\rho_3 = -2$ hence the last result. Note also when the third order moments $E \left[ g(\theta_0) g(\theta_0)' g_j(\theta_0) \right] = 0 \ (j = 1, ..., l)$ then $\text{Bias} \left( \hat{\theta}_{GEL} \right) = \text{Bias} \left( \hat{\theta}_{EL} \right)$.

**Theorem 4** Under the same assumption of Theorem 1

$$\Xi - \Xi_{EL} \text{ is p.s.d.}$$

where $\Xi$ is the higher-order variance of any bias corrected GEL or efficient GMM estimator and $\Xi_{EL}$ is that of EL.

**Proof.** See Theorem 6.1 of Newey & Smith (2004) $\blacksquare$

**Remark 4** This efficiency property of EL will be shared by any estimator for which $\rho_3 = -3$ and $\rho_4 = -6$. 

CAEPR mini-course in empirical likelihood methods
We now compare GEL test statistics in terms of local power. We focus as in Bravo (2003) on the CR statistic, and for simplicity consider the just identified case. We consider local alternatives to $H_0 : \theta = \theta_0$ subject to a Pitman drift $H_n : \theta = \theta_0 + \delta/n^{1/2}$ some nonrandom vector $\delta$ such that $0 < \delta'\delta < \infty$. This hypothesis can be equivalently expressed in terms of that is $H_0 : \lambda = 0$ and $H_n : \lambda = 0 + \delta/n^{1/2}$, which effectively corresponds to considering the following moment conditions $E[g(\theta)] = \delta/n^{1/2}$.

Note that we use the so-called index notation and indicate arrays by their elements. Thus, for any index $1 \leq r_j \leq k$ ($j = 1, 2, \ldots, k$), $a_r$ is an $\mathbb{R}^k$-valued vector, $a_{rs}$ is an $\mathbb{R}^{k \times k}$-valued matrix, etc.

Theorem 5 Assume that (I) $E \|g(\theta_0)\|^{\alpha} < \infty$ for $\alpha \geq 6$, (II)

$$\lim_{\|t\| \to \infty} \left| E \left[ \exp \left( t^{1/2} t' g(\theta_0) \right) \right] \right| < 1.$$ 

Then the second-order power function of the $\text{CR}(\theta_0)$ statistic has the valid Edgeworth
expansion

\[
\Pr \left( C R \left( \Omega_0^{1/2} \delta / n^{1/2} \right) \geq c_\alpha \right) = G_{k,\tau}^- (c_\alpha) + \left[ P_0 (c_\alpha, \delta) + P_\gamma (c_\alpha, \delta) \right] / n^{1/2},
\]

where

\[
P_0 (c_\alpha, \delta) = \sum_{r,s,t=1}^k \alpha_{rst} \delta_r \delta_s \delta_t \left\{ 2 \nabla G_{k,\tau}^- (c_\alpha) + \nabla^2 G_{k,\tau}^- (c_\alpha) \right\} / 6,
\]

\[
P_\gamma (c_\alpha, \delta) = \gamma \left[ 3 \sum_{r,s=1}^k \alpha_{rss} \delta_r \nabla G_{k+2,\tau}^- (c_\alpha) + \sum_{r,s,t=1}^k \alpha_{rst} \delta_r \delta_s \delta_t \nabla G_{k+4,\tau}^- (c_\alpha) \right] / 6
\]

\[
\alpha_{rst} = E \left[ \sum_{r_1,s_1,t_1=1}^k \Omega_{0rr_1}^{-1/2} g_{r_1} (\theta_0) \Omega_{0ss_1}^{-1/2} g_{s_1} (\theta_0) \Omega_{0tt_1}^{-1/2} g_{t_1} (\theta_0) \right],
\]

\(G_{\cdot,\tau} (\cdot)\) is the cumulative distribution of a noncentral \(\chi^2\) with noncentrality parameter \(\tau = \delta' \delta\), \(\nabla^k G_{q,\tau} (\cdot) = \sum_{j=0}^k (-1)^j \binom{k}{j} G_{q+2(k-j),\tau} (\cdot)\), and \(c_\alpha = \Pr (\chi_k^2 \geq c_\alpha) = \alpha\).
Proof. Calculations show that the signed squared root of \( CR\left(\Omega_0^{1/2}\delta'/n^{1/2}\right) \) is the \( k \)-dimensional vector \( CR_r \) with components

\[
CR_r = A_r + \delta_r/n^{1/2} + \sum_{s,t=1}^{q} \left( (\gamma + 2) \alpha_{rst} \left( A_t + \delta_t/n^{1/2} \right) /3 - A_{rs} \right) \left( A_s + \delta_s/n^{1/2} \right) /2,
\]

\[
A_{r_1...r_v} = \sum_{i=1}^{n} \sum_{s_1,...,s_v=1}^{k} \left( \Omega_{0r_1s_1}^{-1/2} g_{is_1}(\theta_0) \right)
\]

and cumulants \( k_{r_1,...,r_v} \) given by

\[
\begin{align*}
  k_{r} &= \delta_r + k_{r}^{11}/n^{1/2} + O\left(n^{-1}\right), \quad k_{r,s} = \delta_{rs} + k_{r,s}^{21}/n^{1/2} + O\left(n^{-1}\right), \\
  k_{r,s,t} &= k_{r,s,t}^{31}/n^{1/2} + O\left(n^{-1}\right), \quad k_{r_1,...,r_v} = O\left(n^{-1}\right) \quad v \geq 4,
\end{align*}
\]
where

\[ k_{r}^{11} = \left[ (\gamma - 1) \sum_{s=1}^{k} \alpha_{r s s} + (\gamma + 2) \sum_{s, t=1}^{k} \alpha_{r s t} \gamma_{s} \gamma_{t} \right] / 6, \]

\[ k_{r, s}^{21} = (2\gamma + 1) \sum_{t=1}^{k} \alpha_{r s t} \gamma_{t} / 3, \quad k_{r, s, t}^{31} = \gamma \alpha_{r s t}. \]

Let \( h_{r_1...r_v} \) denote the \( v \)th-order multivariate Hermite tensor associated with \( \phi(\delta, I) \) - the density of the \( k \)-variate normal distribution with mean \( \delta \) and identity covariance matrix. The second-order Edgeworth series is

\[ H \{ \phi(\gamma, I) \} := \phi(\gamma, I) \left[ 1 + \left( \sum_{r=1}^{q} k_{r}^{11} h_{r} + \frac{1}{2} \sum_{r, s=1}^{q} k_{r, s}^{21} h_{r s} + \frac{1}{3!} \sum_{r, s, t=1}^{q} k_{r, s, t}^{31} h_{r s t} \right) / n^{1/2} \right], \]

where the \( k \)'s are the cumulants defined in (6). A formal Edgeworth expansion for the
local power for \( CR \left( \Omega_0^{1/2} \delta / n^{1/2} \right) \) can be derived by computing the multiple integral

\[
Pr \left( n^{1/2} CR \left( \Omega_0^{1/2} \delta / n^{1/2} \right) < c_\alpha^{1/2} \right) = \int_{-\infty}^{c_\alpha^{1/2}} H \{ \phi (\gamma, I) \} \, dw.
\]

Calculations show that the required Edgeworth expansion is given by a linear combination of terms having the form:

\[
Pr \left( CR \left( \Omega_0^{1/2} \delta / n^{1/2} \right) \geq c_\alpha \right) = \int_{S_{c_\alpha}} \prod_{r=1}^{n} \left\{ 1 + \sum_{R_v} k_{R_v} \partial_0^v \exp (w't) \right\} \phi (\delta_r, 1) \, dw_r,
\]

where \( S_{c_\alpha} = \left\{ w \in \mathbb{R}^k : w'w \geq c_\alpha \right\} \), \( \partial_0^v (\cdot) = \partial^v (\cdot) / \partial t_{r_1} \cdots \partial t_{r_v} \bigg|_{t_r=0} \), \( t \) is an \( \mathbb{R}^k \)-valued vector of auxiliary parameters, \( k_{R_v} \) are defined in (6), and the summation is over any index in the sets of indices \( R_v = (r_1, \ldots, r_v) \) for \( v = 1, 2, 3 \). The result follows by calculating the integral and some simplifications.
Theorem 5 show that no test statistic in the CR family dominates in terms of second order power.
The following theorem shows however that it is still obtain a meaningful comparison in terms of local maximinity as defined in Mukerjee (1994). Let $\widetilde{P}(c_\alpha, \tau) := \min P(c_\alpha, \gamma)$, where the minimum is over $\delta \in S_\tau = \left( \delta \in \mathbb{R}^k : \delta'\delta = \tau \right)$.

**Theorem 6** Let $\widetilde{P}_0(c_\alpha, \tau)$ denote $\widetilde{P}(c_\alpha, \tau)$ for EL. Then, under the same of Theorem 5, there exists a positive $\tau_0$ such that

$$\widetilde{P}_0(c_\alpha, \tau) \geq \widetilde{P}_\gamma(c_\alpha, \tau) \quad \text{for } \gamma \neq 0,$$

whenever $0 < \tau < \tau_0$.

**Proof.** See Bravo (2003) ■
4 GEL for weakly identified models

Recently there has been considerable interest to study the asymptotic properties of moment conditions models where the parameter vector is weakly identified, that is when the moment conditions are nearly uninformative for the parameter vector. One prominent example of this situation is when a set of instrumental variables are weakly correlated with the set of endogenous variables\(^2\).

In this situation it is well-known that the standard normal and chi-squared approximations to econometric estimators and test statistics can be poor.

Guggenberger & Smith (2005) suggested to use GEL and proposed a test statistic which is asymptotically pivotal regardless of the strength of identification. This contrast with classical statistics such as the likelihood ratio and the Wald statistic.

Let \( \theta_0 = [\alpha'_0, \beta'_0]' \in \mathbb{R}^{p+q} \) denote the unknown vector of parameters, let \( \theta = [\alpha', \beta'_0]' \) and assume that \( E[g(z_i, \theta)] \) may be close to zero, that is \( \alpha \) is only weakly identified.

To be more specific assume as in Stock & Wright (2000) and Guggenberger & Smith (2005)

\begin{equation}
\text{WID (1)} (i) \ E[\hat{g}(\theta)] = n^{-1/2}m_{1n}(\theta) + m_2(\beta) \text{ where } m_j(\cdot) \ (j = 1, 2) \text{ are continuous functions}
\end{equation}

\(^2\) This is why in the recent econometric literature weak identification is often called “weak instruments”.
such that $\sup_{\theta \in \Theta} \| m_{1n} (\theta) - m_1 (\theta) \| = o (1), m_{1n} (\theta_0) = 0, m_2 (\beta) = 0$ iff $\beta = \beta_0$, (ii) $m_2 (\cdot)$ is a continuously differentiable function in a neighbourhood $\mathcal{N}_{\beta_0}$ of $\beta_0$, (iii) $\text{rank} \left[ M_{\beta_0} \right] = q$ where $M_{\beta_0} = \partial m_2 (\beta_0) / \partial \beta'$.

WID (2) (i) $\Omega (\cdot)$ is continuous on $A \times \beta_0$ and bounded on $\Theta$ (ii) $\Omega (\cdot)$ is nonsingular for all $\theta \in A \times \beta_0$, (iii) $\sup_{\theta \in \Theta} \left\| \hat{\Omega} (\theta) - \Omega (\theta) \right\| = o_p (1)$, (iv) $\sup_{\theta \in A \times \beta_0} \sum_{i=1}^n \| g (z_i, \theta) g (z_i, \theta)' \| / n = O_p (1)$, (iv) $n^{1/2} (\hat{g} (\theta) - E [\hat{g} (\theta)]) \overset{w}{\to} W (\theta)$ in the space of all uniformly continuous with respect to the uniform norm bounded $\mathbb{R}^{p+q}$-valued functions $BL^\infty$, where $W (\theta)$ is a centred Gaussian process with covariance function $\text{COV} (W (\theta_1), W (\theta_2)) = \Delta (\theta_1, \theta_2)$, (iv) $\sup_{\theta \in A \times \beta_0} \| W (\theta) \| = O_p (1)$.

**Theorem 7 (Guggenberger and Smith (2005))** Assume that (I), (III) of Theorem 1 and WID(1)-(2) hold. Then

$$2n \left( \hat{P}_\rho \left( \hat{\lambda}, \theta \right) - \hat{P}_\rho (0, \theta) \right) \overset{d}{\to} \chi^2_1 (c),$$

where $c = m_1 (\theta)' \Delta (\theta)^{-1} m_1 (\theta)$. 

CAEPR mini-course in empirical likelihood methods
Proof. By Guggenberger & Smith (2005) $\hat{\alpha}$ is inconsistent and $\hat{\beta}$ is $n^{1/2}$ consistent. Their results show also that $\hat{\lambda} = -\hat{\Omega}(\theta)^{-1}\hat{g}(\theta) + o_p(1),$ so that by the same expansion of $\hat{P}_\rho(\hat{\lambda}, \theta)$ about $\theta$ as that in Theorem 1 it follows that

$$\hat{P}_\rho(\hat{\lambda}, \theta) = n\hat{g}(\theta)'\hat{\Omega}(\theta)^{-1}\hat{g}(\theta) + o_p(1)$$

and since $\hat{\Omega}(\theta)^{-1} \xrightarrow{p} \Delta(\theta)^{-1}$ and $\hat{g}(\theta) \xrightarrow{d} N(m_1(\theta), \Delta(\theta))$ the conclusion follows by CMT.

Remark 5 Guggenberger & Smith (2005) proposed two additional tests statistics that are direct by-products of GEL estimation: a score type statistic

$$S_\rho = n\hat{\lambda}'G_\rho(\theta) \left(G_\rho(\theta)'\Delta(\theta)^{-1}G_\rho(\theta)\right)^{-1}G_\rho(\theta)'\hat{\lambda},$$

and a Lagrange multiplier type statistic

$$LM_\rho = n\hat{g}(\theta)'\Delta(\theta)^{-1}G_\rho(\theta) \left(G_\rho(\theta)'\Delta(\theta)^{-1}G_\rho(\theta)\right)^{-1}G_\rho(\theta)'\Delta(\theta)^{-1}\hat{g}(\theta)$$
where \( G(\theta) = \sum_{i=1}^{n} \rho \left( \hat{\lambda} g_i(\theta) \right) \partial g_i(\theta) / \partial \theta \). Both statistics are shown to converge in distribution to \( (\xi_1(\theta) + \xi_2)' (\xi_1(\theta) + \xi_2) \) where \( \xi_2 \sim N(0, I) \) and \( \xi_1(\theta) \) is a random vector depending on \( \theta \).

Note also that for \( \theta_0 \)

\[
2n \left( \hat{P}_\rho \left( \hat{\lambda}, \theta_0 \right) - \hat{P}_\rho \left( 0, \theta_0 \right) \right) \xrightarrow{d} \chi^2_l,
\]

\[
S_\rho, LM_\rho \xrightarrow{d} \chi^2_k.
\]
5 Computation of GEL

- Unless the hypothesis of interest is simple, GEL inference requires solving a saddle-point problem, which in some cases is not that trivial. Here we briefly consider the two GEL methods most used in practice: EL and ET. Recall that the GEL estimator $\hat{\theta}$ is the solution to

$$
\hat{P}_\rho \left( \hat{\theta}, \lambda \right) := \min_{\theta \in \Theta} \sup_{\lambda \in \Lambda_n(\theta)} \hat{P}_\rho (\theta, \lambda)
$$

There are two main approaches to solve (7).
- The GEL estimator can be computed using a nested algorithm that uses a literal interpretation of the saddlepoint property of the estimator. The inner stage is to maximise $\hat{P}_\rho (\theta, \lambda)$ over $\lambda$ for a fixed initial value of $\theta$. Let $\lambda (\theta)$ be the maximising value of $\lambda$. The outer stage is to minimise $\hat{P}_\rho (\theta, \lambda (\theta))$ over $\beta$. The inner stage is numerically simple because it requires maximising a concave function over a convex domain. The second one is more difficult and ultimately depends on the functional form of the moment indicator $g$:

For smooth $g$ quasi-Newton methods can be very effective. Alternatively one can use NPSOL, as for example suggested by Owen (2001) for EL.
For nonsmooth $g$ one can use Nelder-Mead, or simulated annealing.

- The GEL estimator can be computed using a nonnested algorithm for smooth $g$, and use the Newton algorithm (or quasi Newton) to solve simultaneously the FOCs

\[
\begin{bmatrix}
\frac{\partial \hat{P}_\rho(\theta, \lambda)}{\partial \lambda} \\
\frac{\partial \hat{P}_\rho(\theta, \lambda)}{\partial \theta}
\end{bmatrix} = 0
\]

The advantage of this method is that it is faster than the nested one, and for good starting values the convergence is typically quadratic. However for starting values not close enough we typically converges to a saddlepoint quite far away from the true values of the parameters.
It is important to note that for GEL estimators belonging to the ECR family the constraint 
\( \pi_i \geq 0 \) (which is relevant for \( \gamma \geq 0 \)) can be eliminated by introducing a modified objective function which is still concave in \( \lambda \) but prevents its argument from becoming too small. For example for EL one could use (Owen 2001) \[ \sum_{t=1}^{n} \log^* \left( 1 - \lambda' g(\theta) \right) \] where

\[
\log^*(v) = \begin{cases} 
\log(v) & \text{if } v \geq \varepsilon \\
\log(\varepsilon) - 1.5 + 2v/\varepsilon - v^2/(2\varepsilon^2) & \text{if } v < \varepsilon
\end{cases}
\]

for some small \( \varepsilon > 0 \). For example \( \varepsilon = 1/n \). Note that neither ET nor CU have this problem.
- Starting values for both methods can be based on GMM for \( \theta_0 \) and a small value for \( \lambda \).
5.1 Hybrid methods

We consider methods that can be used to avoid the computation of the saddlepoint.

- **One-step GEL.**
  
  This method is based on
  \[
  \hat{P}_\rho \left( \hat{\theta}, \hat{\lambda}^1 \right) : = \min_{\theta \in \Theta} \hat{P}_\rho \left( \theta, \hat{\lambda}^1 \right)
  \]
  
  \[
  \hat{\lambda}^1 = \left( \sum_{i=1}^{n} g_i \left( \hat{\theta}_{GMM} \right) g_i \left( \hat{\theta}_{GMM} \right) \right)^{-1} \sum_{i=1}^{n} g_i \left( \hat{\theta}_{GMM} \right)
  \]

  This is the method used in Kitamura and Tripathi (2003) for the conditional (local) EL estimator

- **Plug-in GEL.**
  
  Suppose that
  \[
  E \left[ g \left( \theta_0 \right) \right] = E \left[ g_1 \left( \theta_{10}, \theta_{20} \right), g_2 \left( \theta_{10}, \theta_{20} \right) \right] = 0
  \]

  and that \( \theta_1 \) is parameter of interest. We can use the second moment condition to estimate the nuisance parameter \( \theta_2 \), and then use GEL on \( g_1 \left( \theta_1, \hat{\theta}_2 \right) \). This method
is particularly appealing when the nuisance parameter is infinite dimensional, as it happens, for example, in semiparametric models.
5.2 Software

We briefly discuss what is available on the web for EL, ET and CU.

1. The basic routine to compute EL ratios for multivariate means can be found at: http://www-stat.stanford.edu/~owen/empirical/. The code el.S. is written in S and works in R as well. It is easy to modify for linear and nonlinear models if you want to use the nonnested algorhitm. There is also another code el.R which does EL for means, quantiles, hazards with censored and truncated data. Can be easily modified to deal with censored regressions.

2. Bruce Hansen’s page http://www.ssc.wisc.edu/~bhansen/progs/progs_gmm.html contains GAUSS and MATLAB codes for computing EL using both the nested and nonnested algorithms.

3. Computation of ET is in general easier than EL. The above mentioned programs can be easily modified for ET.


Imbens, G., Spady, R. & Johnson, P. (1998), ‘Information theoretic approaches to infer-
ence in moment condition models’, *Econometrica* 66, 333–357.


