Empirical likelihood methods with applications to econometrics

Francesco Bravo

October 2007

1Prepared for the Center for Applied Economics and Policy Research, Department of Economics, Indiana University
## Contents

1 Introduction ........................................... 1

2 Empirical likelihood for moment conditions models ........ 5  
   2.1 Empirical likelihood for exactly identified moment conditions models ........................................ 5  
   2.2 Empirical likelihood for overidentified moment condition models ........................................... 9  
   2.3 Higher order asymptotic theory (I) ........................................ 14  
      2.3.1 Empirical likelihood and saddlepoint approximation ........................................... 15  
      2.3.2 Bartlett corrections ........................................ 15  

3 Generalised empirical likelihood and empirical discrepancies 19  
   3.1 Higher order asymptotic theory (II) ........................................ 24  
   3.2 Generalised empirical likelihood for weakly identified models ........................................... 27  

4 Implied probabilities, efficient bootstrap and auxiliary information 29  
   4.1 Asymptotic properties of GEL implied probabilities ........................................ 29  
      4.1.1 Appendix ........................................ 32  
   4.2 Efficient bootstrap in moment conditions models ........................................ 33  
      4.2.1 Appendix ........................................ 36  
   4.3 Efficient Z-estimation with auxiliary information ........................................ 37  

5 Generalised empirical likelihood and smoothing ............... 41  
   5.1 Density estimation ........................................ 41  
   5.2 Local linear regression ........................................ 43  
   5.3 Other applications ........................................ 44  

6 Generalised empirical likelihood and dependent data ............... 47  
   6.1 (Generalised) Dual likelihood ........................................ 47
6.1.1 GDL for unstable autoregressive processes . . . . . . . . . . . 49
6.2 Blocking and smoothing methods . . . . . . . . . . . . . . . . . 54

7 Other applications 63
7.1 Non differentiable moment indicators . . . . . . . . . . . . . . . 63
7.2 Conditional moment conditions models . . . . . . . . . . . . . . . 65
7.2.1 Applications of Conditional EL . . . . . . . . . . . . . . . . . 73
Preface

Since its introduction as a nonparametric likelihood based alternative to likelihood and bootstrap methods, Owen’s (1988) empirical likelihood has gained increasing popularity among econometricians and statisticians. This set of lecture notes introduces the basic ideas behind empirical likelihood 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. Some of the results are new and presented here for the first time.

Notation

We use the following abbreviations and symbols:

“CLT” for central limit theorem, “CMT” for continuous mapping theorem, “(U)LLN” for (uniform) law of large numbers,

“w.p.a. 1” for with probability approaching 1, “a.s.” for almost surely, “p.d.” for positive definite “$\mathbb{P}$”, “$\mathbb{D}$” for convergence in probability and in distribution,

“T”, “M”, “CS”, “J”, for Triangle, Markov, Cauchy-Schwarz and Jensen inequalities,

$\|\cdot\|$ is the Euclidean norm, $\sigma_{\min}()$, $\sigma_{\max}()$ are the minimum and maximum eigenvalue operators,

$C$ is a generic positive constant, $I \{\cdot\}$ is the indicator function.
Chapter 1

Introduction

Owen’s (1988) empirical likelihood (EL henceforth) 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:

1. The shape of confidence regions is data determined (Owen 2001),

2. When constraints are known to hold among the parameters of interest, they can be imposed numerically (Owen 2001),

3. A Bartlett correction applies (DiCiccio, Hall & Romano 1991),

4. All points in confidence regions obey range restrictions: variances are non-negative, probabilities are between [0, 1], correlations are between [-1, 1] (Owen 2001),

5. Transformation invariant and internal studentisation (Owen 2001),

6. Second-order maximinity (Bravo 2003),

7. It is optimal in the Generalized Neyman-Pearson Lemma (GNP) sense (Kitamura 2001).
Further properties of EL (in terms of bias and higher order efficiency for estimators) are discussed in Section 1 of Chapter 3. Properties 3 and 6 are discussed in Section 3 of Chapter 2 and Section 1 of Chapter 3, respectively. The GNP optimality of EL implies that under certain conditions EL tests are uniformly more powerful in a Large Deviations sense\footnote{Further connections between Large Deviations and EL can be found in Kitamura (2006).}

Perhaps the best way to introduce EL is by a direct comparison with 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 \(\theta\), and let \(\hat{\theta} = \arg \max_{\theta} L(\theta)\) denote the maximum likelihood estimator. Suppose that we are interested to test the hypothesis \(H_0 : \theta = \theta_0\). Let

\[
R(\theta_0) = L(\theta_0) / L(\hat{\theta})
\]

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 \(\pi_i\) 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 \(\pi_i\)s 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 \(\hat{F}(z) = \sum_{i=1}^{n} I\{z_i \leq z\} / n\) with probability \(\pi_i = 1/n\). Thus in analogy to ordinary likelihood methods we can define a nonparametric likelihood ratio as

\[
R(F)^2 = \prod_{i=1}^{n} \pi_i / \prod_{i=1}^{n} 1/n = \prod_{i=1}^{n} n \pi_i.
\]

\footnote{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.}
\( R(F) \) is used by EL to obtain inferences about an unknown parameter \( \theta(F) \) by finding a \( \pi \), consistent with \( \theta(F) \). This is what EL does in practice.

Interestingly 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). As noted by Owen (2001) the 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 \( \hat{S}(z) \) to define the ratio \( R(S) = L(S) / L(\hat{S}) \).
Chapter 2

Empirical likelihood for moment conditions models

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

\[
E[g(z_i, \theta_0)] = 0,
\]

for a unique unknown \(\theta_0\), where the expectation \(E\) is with respect \(F\). In this chapter we investigate how EL can be used to obtain inferences about \(\theta_0\), and more generally about \(h(\theta_0)\) for some nonlinear (differentiable) function \(h(\cdot)\) assuming that \(g(\cdot)\) is differentiable. The nondifferentiable case will be discussed in Section 1 of Chapter 7.

2.1 Empirical likelihood for exactly identified moment conditions models

In this section we assume that the dimension of the moment indicator \(g(\cdot)\) is the same as that of the unknown parameter vector (i.e. the model is exactly identified).

**Example 1** (Linear regression) Let \(z_i = [y_i, x_i']\), and let \(y_i = x_i'\theta_0 + \varepsilon_i\) where \(\varepsilon_i\) is an unobservable error with \(E(\varepsilon_i) = 0\). Then the orthogonality condition between the error and the regressors defines a moment condition model with

\[
E[x_i (y_i - x_i'\theta_0)] = 0.
\]
Example 2 (Quasi-ML) Let \( l(z_i, \theta) \) denote the quasi-loglikelihood function for \( \theta \), and let \( s(z_i, \theta) = \partial l(z_i, \theta) / \partial \theta \) denote the quasi-score vector. Then the property that \( E[s(z_i, \theta)] = 0 \) defines a moment condition model with

\[
e[s(z_i, \theta)] = 0.
\]

Let \( g(z_i, \theta) = g_i(\theta), \sum_{i=1}^{n} g(z_i, \theta) / n = \hat{g}(\theta) \), and analogously \( \partial g(z_i, \theta) / \partial \theta = G_i(\theta), \sum_{i=1}^{n} (\partial g(z_i, \theta) / \partial \theta) / n = \hat{G}_i \). Typically \( \theta_0 \) is estimated using the so-called analogy principle in which the expectation is replaced by its sample analogue, and the estimator is defined as

\[
\left\| \hat{g} (\hat{\theta}) \right\| = 0.
\]

We shall call this estimator a Z estimator\(^2\) Consistency of \( \hat{\theta} \) and asymptotic normality of \( n^{1/2} (\hat{\theta} - \theta_0) \) can be established under mild regularity conditions, as the following theorem shows.

**Theorem 3** Assume that (I) the parameter space \( \Theta \) is a compact set, (II) for all \( \zeta > 0 \inf_{||\theta - \theta_0|| > \zeta} ||Eg(\theta)|| \geq \varepsilon (\zeta) > 0 \), (III) \( \sup_{\theta \in \Theta} \left\| \hat{g}(\theta) - Eg(\theta) \right\| = o_p(1) \). Then \( \hat{\theta} \xrightarrow{P} \theta_0 \). Assume further that (IV) \( \theta_0 \in \text{int} \{\Theta\} \), (V) \( \sup_{\theta \in N_0} \left\| \hat{G}(\theta) - E[G(\theta)] \right\| = o_p(1) \), (VI) \( E(G_0) \) is nonsingular where \( G_0 = G(\theta_0) \), (VII) \( n^{1/2} \hat{g}(\theta_0) \xrightarrow{d} N(0, \Omega_0) \) where \( \Omega_0 = E[g(\theta_0)g(\theta_0)'] \). Then

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

**Proof.** Note that

\[

\left\| Eg (\hat{\theta}) \right\| \leq \sup_{\theta \in \Theta} \left\| \hat{g}(\theta) - Eg(\theta) \right\| + \left\| \hat{g} (\hat{\theta}) \right\| \\
\leq \sup_{\theta \in \Theta} \left\| \hat{g}(\theta) - Eg(\theta) \right\| = o_p(1) + ||Eg(\theta_0)|| = o_p(1)
\]

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

Inference about \( \theta \) 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 \( z_i \)).

Here we consider how EL method can be used to test the simple hypothesis \( H_0 : \theta = \theta_0 \). Recall from the Introduction that EL uses a multinomial supported

---

\(^1\)In the sense of White (1982)

\(^2\)This estimator is also called moment estimator and occasionally an M-estimator.
2.1. EMPIRICAL LIKELIHOOD FOR EXACTLY IDENTIFIED MOMENT CONDITIONS MODELS

on the sample. In the case of moment condition models the multinomial is

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

(2.2)

The solution to (2.2) 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

\[ \partial \mathcal{L}(\pi_i, \lambda, \gamma) / \partial p_i = 1 / \pi_i - \hat{\lambda}' g_i (\theta_0) + \hat{\gamma} = 0, \]
\[ \partial \mathcal{L}(\pi_i, \lambda, \gamma) / \partial \lambda = \sum_{i=1}^{n} \pi_i g_i (\theta_0) = 0, \]
\[ \partial \mathcal{L}(\pi_i, \lambda, \gamma) / \partial \gamma = \sum_{i=1}^{n} \pi_i - 1 = 0. \]

Summing over the \( i \) the first line and multiplying by \( \pi_i \) yields

\[ \sum_{i=1}^{n} 1 / \pi_i - \hat{\lambda}' \sum_{i=1}^{n} g_i (\theta_0) + \sum_{i=1}^{n} \hat{\gamma} = 0 \Rightarrow \sum_{i=1}^{n} (1 + \hat{\gamma} \pi_i) = 0 \]
\[ \hat{\gamma} = -n. \]

Replacing this to the first line and multiplying both sides by \( \pi_i \) gives

\[ \hat{\pi}_i^{-1} = n \left( 1 + \hat{\lambda}' g_i (\theta_0) \right). \]  

(2.3)

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

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

Let

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

denote the (log) EL ratio test statistic. The following theorem is the non-parametric likelihood version of Wilks’ theorem adapted from Owen (1988) and Owen (1990) to moment conditions models.
CHAPTER 2. EMPIRICAL LIKELIHOOD FOR MOMENT CONDITIONS MODELS

**Theorem 4** Assume that (I) \(0 \in \text{ch} \{g_1(\theta_0), \ldots, g_n(\theta_0)\}\), (II) \(\mathbb{E}[\|g(\theta_0)\|^2] \leq \infty\), (III) \(\mathbb{E}[g(\theta_0)g(\theta_0)'] = \Omega_0\) p.d.. Then under \(H_0\)

\[ W(\theta_0) \overset{d}{\to} \chi_k^2. \]

**Proof.** First we establish the convergence rate for the Lagrange multiplier \(\hat{\lambda}\).

By definition \(\hat{\lambda} = \eta\rho\) where \(\|\eta\| = 1\) and \(\rho > 0\). Then (2.4) can be written as

\[ 0 = \hat{g}(\theta_0) - \sum_{i=1}^n g_i(\theta_0) g_i(\theta_0)' \eta\rho/n (1 + \eta' \rho g_i(\theta_0)). \]

Multiplying the right hand side by \(\eta'\) and rearranging

\[ 0 = \eta' \sum_{i=1}^n g_i(\theta_0) (1 + \eta' \rho g_i(\theta_0)) / n - \eta' \hat{\Omega}(\theta_0) \eta\rho/n \]

and note that \(\sigma_{\min}(\Omega_0) \leq \eta' \hat{\Omega}(\theta_0) \eta \leq \sigma_{\max}(\Omega_0)\) w.p.a.1, and that \(\|\hat{g}(\theta_0)\| = O_p\left(n^{-1/2}\right)\) by CLT and \(\max_i \|g_i(\theta_0)\| = o_p\left(n^{1/2}\right)\) by Borel-Cantelli lemma so that

\[ 0 \leq \|\eta\| \|\hat{g}(\theta_0)\| \left(1 + \|\eta\| \rho \max_i \|g_i(\theta_0)\|\right) - \sigma_{\min}(\Omega_0) \rho \]

\[ 0 \leq O_p\left(n^{-1/2}\right) \left(1 + \rho o_p\left(n^{1/2}\right)\right) - \sigma_{\min}(\Omega_0) \rho \]

which shows that \(\rho = O_p\left(n^{-1/2}\right)\) i.e. \(\hat{\lambda} = O_p\left(n^{-1/2}\right)\). Next we find a stochastic approximation for \(\hat{\lambda}\). Using again (2.4) and noting that \(\max_i |\hat{\lambda}' g_i(\theta_0)| = o_p(1)\) we have

\[ 0 = \hat{g}(\theta_0) - \sum_{i=1}^n g_i(\theta_0) g_i(\theta_0)' \hat{\lambda} / n \left(1 + \hat{\lambda}' g_i(\theta_0)\right) \Rightarrow \]

\[ \hat{\lambda} = \frac{\hat{\Omega}(\theta_0)^{-1} \hat{g}(\theta_0) + o_p(1) = \Omega_0^{-1} \hat{g}(\theta_0) + o_p(1).} {2} \]

Finally by Taylor expansion

\[ W(\theta_0) = 2 \sum_{i=1}^n \log \left(1 + \hat{\lambda}' g_i(\theta_0)\right) = 2\hat{\lambda}' \sum_{i=1}^n g_i(\theta_0) - \]

\[ 2 \lambda' \sum_{i=1}^n g_i(\theta_0) g_i(\theta_0)' \hat{\lambda} + o_p\left(\sum_{i=1}^n (\hat{\lambda}' g_i(\theta_0))^3\right). \]

\(^3\)The convex hull condition (see Footnote 2) 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.
Replacing (2.5) in (2.6) and noting that
\[ \sum_{i=1}^{n} (\hat{\lambda}' g_i (\theta_0)) \leq \max_i |\hat{\lambda}' g_i (\theta_0)| \hat{\lambda} \sum_{i=1}^{n} g_i (\theta_0) g_i (\theta_0)' \hat{\lambda} = o_p (1) \]
we get
\[ W (\theta_0) = n \hat{\theta} (\theta_0)' \Omega_0^{-1} \hat{\theta} (\theta_0) + o_p (1) \]
from which the conclusion follows by CMT.

**Remark 5** 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 \( \lambda \) 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 \( \lambda \) is the Lagrange multiplier associated with the sample moment condition \( \sum_{i=1}^{n} \pi_i g_i (\theta_0) = 0 \). The resulting test statistic is given by twice the maximised function \( \sum_{i=1}^{n} \log (1 + \lambda g_i (\theta_0)) \) with respect to \( \lambda \). This is the idea behind a number of artificial likelihoods including Mykland’s (1995) dual likelihood (discussed in some detail in Chapter 6), Smith’s (1997) generalised empirical likelihood (discussed in some detail in Chapter 3), and in Chesher & Smith’s (1997) augmented densities for specification testing.

### 2.2 Empirical likelihood for overidentified moment condition models

In this section we concentrate on the so-called overidentified moment conditions models, that is models where the number of equations defining implicitly the parameters is bigger than the number of parameters itself. Assume that the dimension of \( g (\cdot) \) is \( l \) \( (l > k) \) while the dimension of \( \theta \) is still \( k \). As before
\[ E [g_i (\theta_0)] = 0 \]
for a unique \( \theta_0 \).

**Example 6** (Generalised instrumental variable regression). Let \( z_i = [y_i, x_i', w_i']' \), and let \( y_i = x_i' \theta_0 + \varepsilon_i \) where \( \varepsilon_i \) 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 \( \text{rank} (E [w_i x_i']) = k \) 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 \( \theta_0 \) is estimated using Hansen’s (1982) generalised method of moment (GMM) method, and the GMM (or generalised Z) estimator is defined as

\[
\| \hat{g} \left( \hat{\theta}_{GMM} \right) \|_{\hat{W}} = \inf_{\theta \in \Theta} \| \hat{g} (\theta) \|_{\hat{W}}
\]

where \( \hat{W} \) 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

\[
n^{1/2} \left( \hat{\theta}_{GMM} - \theta_0 \right) \overset{d}{\rightarrow} N \left( 0, (G_0' W G_0)^{-1} G_0' W \Omega_0 W G_0 (G_0' W G_0)^{-1} \right)
\]

(2.8)

where \( W = p \lim (\hat{W}) \). Note that (2.8) 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 \( \Omega_0^{-1} \). In this case we obtain the so-called efficient GMM estimator, that is

\[
n^{1/2} \left( \hat{\theta}_{GMM} - \theta_0 \right) \overset{d}{\rightarrow} N \left( 0, \left( G_0' \Omega_0^{-1} G_0 \right)^{-1} \right).
\]

(2.9)

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

\[
J_n \left( \hat{\theta}_{GMM} \right) = n \hat{g} \left( \hat{\theta}_{GMM} \right)' \hat{\Omega} \left( \hat{\theta}_{GMM} \right)^{-1} \hat{g} \left( \hat{\theta}_{GMM} \right) \overset{d}{\rightarrow} \chi^2_{1-p}.
\]

(2.10)

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 \( \text{rank} (H(\theta)) = p \) where \( H(\theta) = \partial H(\theta) / \partial \theta' \). Let

\[
\| \hat{g} \left( \hat{\theta}_{GMM}^c \right) \|_{\hat{W}} = \inf_{\theta \in \Theta} \| \hat{g} (\theta) \|_{\hat{W}} \quad \text{s.t.} \quad h(\theta) = 0
\]

denote the constrained GMM estimators, and let

\[
D = J_n \left( \hat{\theta}_{GMM} \right) - J_n \left( \hat{\theta}_{GMM}^c \right)
\]

(2.11)

\[
LM = n \left( \hat{\gamma}^c \right)' \tilde{\Phi} \left( \hat{\theta}_{GMM}^c \right) \hat{\gamma}^c, \quad W = nh \left( \hat{\theta}_{GMM} \right)' \tilde{\Phi} \left( \hat{\theta}_{GMM} \right)^{-1} h \left( \hat{\theta}_{GMM} \right)
\]

\[
\Phi(\cdot) = H(\cdot) \left( \hat{G}(\cdot)' \hat{\Omega}(\cdot)^{-1} \hat{G}(\cdot) \right)^{-1} H(\cdot)'
\]

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

\[
D, LM, W \overset{p}{\rightarrow} \chi^2_p.
\]
2.2. EMPIRICAL LIKELIHOOD FOR OVERIDENTIFIED MOMENT CONDITION MODELS

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 \( \theta_0 \). Thus the estimation process becomes more complicated as the EL estimator \( \hat{\theta} \) is defined as

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

(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 (1997b)) shows that the EL estimator has the same asymptotic distribution as that of the efficient GMM estimator. Let \( \Gamma(\theta, \delta) \) denote an open sphere centred at \( \theta \) with radius \( \delta \).

**Theorem 7** Assume that (I) the parameter space \( \Theta \) is compact, (II) for a small \( \delta > 0 \) \( E \sup_{\theta \in \Gamma(\theta, \delta)} - \log \left( 1 + \lambda' g_i(\theta^*) \right) < \infty \), (III) \( \theta_0 \in \text{int}(\Theta) \), (IV) \( g(z_i, \theta) \) is twice continuously differentiable at \( \theta_0 \), (V) \( \theta_0 \) is p.d. (VI) \( E \sup_{\theta \in \Gamma(\theta, \delta)} \| g(\theta^*) \|^2 + \epsilon < \infty \) for some \( \epsilon > 0 \), \( E \sup_{\theta \in \Gamma(\theta, \delta)} \| G(\theta^*) \|^2 < \infty \), \( E \sup_{\theta \in \Gamma(\theta, \delta)} \| \partial G_i(\theta^*) / \partial \theta_j \| < \infty \) \( (j = 1, ..., k) \). Then \( \hat{\lambda} \sim 0 \), \( \hat{\theta} \sim 0 \), and

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

(2.12)

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 \( \hat{\lambda} \) follows as in the proof of Theorem 2 which shows that \( \hat{\lambda} = O_p \left( n^{-1/2} \right) \). The proof of consistency of \( \hat{\theta} \) 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 \( \theta_0 \) the sample objective function is bounded away from the maximum the population objective function achieved at \( \theta_0 \). Second that the maximum of the sample objective function is not smaller than its value at \( \theta_0 \). Since the latter converges to its expectation it follows that the maximum belongs to the arbitrary neighbourhood containing \( \theta_0 \), which proves consistency. To be specific note that since the “optimal” \( \lambda \) is actually 0 we have

\[
E - \log \left( 1 + \lambda' g_i(\theta) \right) \leq 0
\]

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

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

(2.14)
Because the parameter space \( \Theta \) is compact we can cover the set \( \Theta (\delta) = \Theta - \Gamma (\theta_0, \delta) \) with \( \Gamma (\theta_j, \delta) \) \((j = 1, \ldots, h)\) so that by (2.14) and \( H_j > 0 \)

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

whence by LLN for all \( j \)

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

for \( H = \min_j H_j \). At the same time note that by (2.13)

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

which shows that \( \hat{\theta} \notin \Theta (\delta) \) w.p.a.1 or \( \Pr (\hat{\theta} \in \Gamma (\theta_0, \delta)) \geq 1 - \varepsilon \), so that the consistency follows since \( \delta \) is arbitrary. The asymptotic normality follows by mean value expansion of the first order conditions for \( \hat{\lambda} \) and \( \hat{\theta} \). To be specific note that by consistency \([0', 0'] = \left[ \frac{\partial W (\hat{\lambda}, \hat{\theta}) / \partial \lambda', \partial W (\hat{\lambda}, \hat{\theta}) / \partial \theta' \right]' \) w.p.a.1 where

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

Then by ULLN

\[
\begin{bmatrix}
0 \\
0
\end{bmatrix} = n^{1/2} \left[ \frac{\partial W (0, \theta_0) / \partial \lambda}{\partial W (0, \theta_0) / \partial \theta} \right] + n^{-1} \left[ \begin{array}{c}
\frac{\partial^2 W (\hat{\lambda}, \hat{\theta}) / \partial \lambda \partial \lambda'}{\partial^2 W (\hat{\lambda}, \hat{\theta}) / \partial \theta \partial \lambda'} \\
\frac{\partial^2 W (\hat{\lambda}, \hat{\theta}) / \partial \theta \partial \lambda'}{\partial^2 W (\hat{\lambda}, \hat{\theta}) / \partial \theta \partial \theta'}
\end{array} \right] n^{1/2} \begin{bmatrix}
\hat{\lambda} \\
\hat{\theta} - \theta_0
\end{bmatrix} + \begin{bmatrix}
\Omega_0 \\
G_0
\end{bmatrix} n^{1/2} \begin{bmatrix}
\hat{\lambda} \\
\hat{\theta} - \theta_0
\end{bmatrix},
\]

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

**Remark 8** (2.12) shows an important aspect of EL method, that is that the estimators \( \hat{\lambda} \) and \( \hat{\theta} - \theta_0 \) 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 \( \theta \). We first consider the EL analogue of Hansen’s (1982) \( J \) general test for misspecification (2.10).
2.2. EMPIRICAL LIKELIHOOD FOR OVERIDENTIFIED MOMENT CONDITION MODELS

Theorem 9 Under the same assumptions of Theorem 3,

\[ 2 \sum_{i=1}^{n} \log \left( 1 + \lambda' g_i \left( \hat{\theta} \right) \right) \xrightarrow{d} \chi^2_{l-k}. \]  

(2.16)

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

\[
2 \sum_{i=1}^{n} \log \left( 1 + \lambda' g_i \left( \hat{\theta} \right) \right) = 2 \sum_{i=1}^{n} \left( \lambda' g_i \left( \hat{\theta} \right) - \left( \lambda' g_i \left( \hat{\theta} \right) \right)^2 / 2 \right) + O \left( \sum_{i=1}^{n} \left( \lambda' g_i \left( \hat{\theta} \right) \right)^3 \right)
\]

where the second line follows by \( n \lambda' \sum_{i=1}^{n} g_i \left( \hat{\theta} \right) = 2 \lambda' \sum_{i=1}^{n} g_i \left( \hat{\theta} \right) + o_p(1) \) (as in the proof of Theorem 2).

Note that (2.16) does not require the explicit estimation of the covariance matrix \( \Omega_0^{-1} \), and this is clearly convenient when such estimation is difficult.

We now consider the general case of EL inference for the same hypothesis \( H_0 : h \left( \theta_0 \right) = 0 \). 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

\[ \hat{\theta}^c = \min_{\theta \in \Theta} \sum_{i=1}^{n} \log \left( 1 + \lambda' g_i \left( \theta \right) \right) \quad \text{s.t.} \quad h \left( \theta \right) = 0 \]

denote the constrained EL estimator. Then we can define in analogy with (2.11)

\[ ELR = 2 \sum_{i=1}^{n} \log \left( 1 + \lambda' g_i \left( \hat{\theta}^c \right) \right) - 2 \sum_{i=1}^{n} \log \left( 1 + \lambda' g_i \left( \hat{\theta} \right) \right) \]

\[ LM = n \left( \gamma' \Phi \left( \hat{\theta} \right) \hat{\gamma} \right), \quad W = nh \left( \Phi \left( \hat{\theta} \right) \Phi \left( \hat{\theta} \right)^{-1} \right) h \left( \hat{\theta} \right), \]

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

Theorem 10 Under the same assumptions of Theorem 2,

\[ ELR, LM, W \xrightarrow{d} \chi^2_{p}. \]

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 \( \mathcal{L}(\theta, \gamma) = -\sum^n_{i=1} \log (1 + \lambda_i \gamma) - \gamma' h(\theta) \), ULLN, CMT and standard manipulations

\[
\begin{bmatrix}
\frac{n^{1/2} \hat{\lambda}}{\sqrt{n^{1/2} \hat{\gamma}^2}} \\
\frac{n^{1/2} (\hat{\theta} - \theta_0)}{\sqrt{n^{1/2} \hat{\gamma}^2}}
\end{bmatrix} = \begin{bmatrix}
A^{-1} \left( I - B (B'A^{-1}B)^{-1} BA^{-1} \right) A^{-1} B (B'A^{-1}B)^{-1} \\
(B'A^{-1}B)^{-1} B' A^{-1} (B'A^{-1}B)^{-1}
\end{bmatrix}^{-1}
\begin{bmatrix}
\frac{n^{1/2} \hat{g}(\theta_0)}{\sqrt{n^{1/2} \hat{\gamma}^2}} \\
0
\end{bmatrix},
\]

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

\[
\begin{bmatrix}
\frac{n^{1/2} \hat{\lambda}}{\sqrt{n^{1/2} \hat{\gamma}^2}} \\
\frac{n^{1/2} (\hat{\theta} - \theta_0)}{\sqrt{n^{1/2} \hat{\gamma}^2}}
\end{bmatrix} \overset{d}{\rightarrow} N \left( \begin{bmatrix}
0 \\
0
\end{bmatrix}, \begin{bmatrix}
\Xi_0 & 0 \\
0 & \left( H_0' \left( G_0' \Omega_0^{-1} G_0 \right)^{-1} H_0' \right)^{-1}
\end{bmatrix} \right)
\]

and the result follows by CMT, noting the asymptotic independence between \( \hat{\gamma}^2 \), \( \hat{\theta} - \theta_0 \) and \( \hat{\lambda}^2 \). Finally the distribution of \( ELR \) can be obtained by expanding both statistics about 0 and since \( \hat{\lambda}^2 = \hat{\Omega} \left( \hat{\theta} \right)^{-1} \hat{g}(\hat{\theta}) \)

\[
ELR = n\hat{g}(\hat{\theta})' \hat{\Omega} \left( \hat{\theta} \right)^{-1} \hat{g}(\hat{\theta}) - n\hat{g}(\hat{\theta})' \hat{\Omega} \left( \hat{\theta} \right)^{-1} \hat{g}(\hat{\theta})
\]

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

**Example 11** As an illustration of the above theorem we consider the same type of hypothesis considered by Qin & Lawless (1994). Let \( \theta_1 \) is a \( p \times 1 \) vector of parameters of interest in \( \theta = [\theta_1', \theta_2']' \) and \( \theta_2 \) is a \( k - p \) vector of nuisance parameters. The hypothesis of interest is \( H_0 : \theta_1 = \theta_{10} \). Then \( \hat{\theta}^* = [\hat{\theta}_{10}, \hat{\theta}_2] \), \( H = \begin{bmatrix} I_p & 0 \end{bmatrix} \) where \( 0 \) is a \( p \times (k - p) \) matrix of zeros.

### 2.3 Higher order asymptotic theory (I)

This section presents some results concerning some higher order asymptotic properties of EL.
2.3. HIGHER ORDER ASYMPTOTIC THEORY (I)

2.3.1 Empirical likelihood and saddlepoint approximation

Let

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

denote the empirical cumulant generating function for the moment indicator

\[ g_i(\theta_0) \]

where (the saddlepoint) \( \hat{\xi}(\theta_0) \) satisfies

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

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

**Proposition 12** Under the same assumptions of Theorem 2, with (II) strengthened to \( E \| g(\theta_0) \|^3 < \infty \). Then

\[
n\hat{K}(\theta_0) = -W(\theta_0) / 2 + \hat{\Gamma} \left( \hat{\delta} \right) / \left( 6n^{1/2} \right) + O \left( n^{-1} \right).
\]

**Proof.** See Monti & Ronchetti (1993).

Relation (2.17) can be used in two ways. First it can be used to obtain a nonparametric approximation to the density of the Z-estimator \( \hat{\theta} \) by replacing \(-W(\theta_0) / 2 + \hat{\Gamma} \left( \hat{\delta} \right) / \left( 6n^{1/2} \right) \) into the empirical saddlepoint approximation at \( \theta_0 \)

\[
\hat{f}(\theta_0) = \left( n/2\pi \right)^{k/2} \left| \hat{K}(\theta_0) \hat{A}(\theta_0) \right|^{-1/2} \left| \hat{A}(\theta_0) \right| \exp \left( n\hat{K}(\theta_0) \right)
\]

where

\[
\hat{A}(\theta_0) = \exp \left( -\hat{K}(\theta_0) \right) \sum_{i=1}^{n} G_i(\theta_0) \exp \left( \hat{\xi}(\theta_0)' g_i(\theta_0) \right) / n
\]

\[
\hat{K}(\theta_0)^\prime = \exp \left( -\hat{K}(\theta_0) \right) \sum_{i} g_i(\theta_0) g_i(\theta_0)' \exp \left( \hat{\xi}(\theta_0)' g_i(\theta_0) \right) / n.
\]

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

2.3.2 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 \( \chi^2 \) distributed test statistic
CHAPTER 2. EMPIRICAL LIKELIHOOD FOR MOMENT CONDITIONS MODELS

$S(\theta_0)$ closer to its nominal size. To be specific $S(\theta_0)$ is said to be Bartlett-correctable if the terms of order $O(n^{-1})$ in the asymptotic distribution of

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

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

In a seminal paper DiCiccio et al. (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

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

where $h_j(\theta_0)$ is the $j$th ($j = 1, ..., k$) component of the vector $h(\theta_0) = \Omega_{0}^{-1/2} g(\theta_0)$, which can be consistently estimated by say $\hat{b}(\hat{\theta})$. Let

$$\hat{W}^B(\theta_0) = W(\theta_0)/\left(1 + \hat{b}(\hat{\theta})/(nk)\right) \quad (2.18)$$

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

**Theorem 13** Assume that (I) $W(\theta_0)$ admits a valid Edgeworth expansion (in the sense of Chandra & Ghosh (1980)), (II) $n^{1/2} \left\| \hat{\theta} - \theta_0 \right\| = O_p(1)$. Then, for some $c_0 \geq 0$

$$\sup_{c \in [c_0, \infty)} \left| \Pr \left( \hat{W}^B(\theta_0) \leq c \right) - \int_0^c \chi^2_k(x) \, dx \right| = O(n^{-2}). \quad (2.19)$$

**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 $\mathbb{R}^b$-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

$$W_r(\theta_0) \sim N\left(\gamma(\theta_0)/n^{1/2}, I + \Gamma(\theta_0)/n\right) + O\left(n^{-3/2}\right), \quad (2.20)$$

where $\gamma(\theta_0)$ and $\Gamma(\theta_0)$ are, respectively, a vector and matrix such that $\gamma(\theta_0)' \gamma(\theta_0) + \text{trace}(\Gamma(\theta_0)) = b(\theta_0)$. 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 proportional to $\chi^2_k(x) [1 + b(\theta_0) x/n] + O(n^{-2})$ 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 $\hat{b}(\mathbf{\hat{\theta}})$ it follows that $\Pr(W^{\hat{B}}(\theta_0) \leq c) = \Pr(W^{B}(\theta_0) \leq c) + O(n^{-3/2})$ where the last term is actually of order $O(n^{-2})$ by the symmetry argument of Barndorff-Nielsen & Hall (1988).

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 over-identified 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 $\Psi$ 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, $\Lambda$ 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 $H_0: \theta = \theta_0$

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

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) shows 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

$$\sup_{c \in [c_0, \infty]} \left| \Pr(W^{\hat{B}}(\theta_0) \leq c) - \int_{0}^{c} \chi^2_k(x) \, dx \right| = O(n^{-2})$$

where $W^{\hat{B}}(\theta_0)$ is as in (2.18) 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.
Chapter 3

Generalised empirical likelihood and 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. Clearly other distances or discrepancies could be considered. For example

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

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

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

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

\(^1\)Note that the two degenerate cases \( \gamma = -1 \) and \( \gamma = 0 \) are handled by taking the limits.
similar Lagrange multiplier argument as that used in (2.2) it is possible to show that the CR test statistic for $H_0 : \theta = \theta_0$ is

$$CR(\theta_0) = \frac{2}{\gamma(\gamma + 1)} \sum_{i=1}^{n} \left[ \left( 1 + \hat{\zeta} + \hat{\lambda}' g_i(\theta_0) \right)^{\gamma/(\gamma + 1)} - 1 \right], \quad (3.1)$$

where Lagrange multipliers $\hat{\zeta} \in \mathbb{R}$ and $\hat{\lambda} \in \mathbb{R}^k$ are determined by the constraints $\sum_{i=1}^{n} \pi_i = 1$ and $\sum_{i=1}^{n} \pi_i g_i(\theta_0) = 0$, respectively. Then $CR(\theta_0) \overset{d}{\to} \chi^2_k$ (Baggerly 1998) All members of the CR family enjoy the following desirable statistical properties: they yield convex confidence regions (at least for a multivariate mean) whose shape is typically data-determined. Furthermore CR regions are transformation invariant, do not require estimation of scale. 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 $\pi_i$ 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 $\pi_i$ 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 is $h(x, x) = 0$, $\partial^r h(x, y)/\partial x^r |_{x=y} = O_p(n^{-r} / k_h)$ for $r = 1, \ldots, 4$, and $\partial^2 h(x, y)/\partial x^2 |_{x=y} \neq 0$, 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 (2.1) (i.e. $H_0 : \theta = \theta_0$) is based on

$$\min_{\pi_i} \sum_{i=1}^{n} k_h d(\pi_i, 1/n) \text{ s.t. } \sum_{i=1}^{n} \pi_i = 1 \text{ and } \sum_{i=1}^{n} \pi_i g(\theta_0) = 0.$$ 

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 \( \Lambda_n(\theta) = \{ \lambda : \lambda' g_i(\theta) \in V, i = 1, ..., n \} \). The GEL estimator is defined as
\[
\hat{\theta} = \min_{\theta \in \Theta} \sup_{\lambda} \hat{P}_\rho(\theta, \lambda),
\]
where
\[
\hat{P}_\rho(\theta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \rho(\lambda' g_i(\theta))/n, \tag{3.2}
\]
that is the GEL estimator is the solution of a saddlepoint problem.

**Example 14** 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 (3.1).

The GEL criterion function (3.2) 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. They propose a likelihood ratio test statistic and show (among other things) that it is Bartlett-correctable. In the context of the moment conditions models considered here however 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 as mentioned in Remark 5 is the dual of \( H_0 : \theta = \theta_0 \). For the CR statistic this is explicitly the case as \( \lambda \) is the Lagrange multiplier which ensures that the moment conditions are satisfied in the sample (cf. (3.1)).

It is interesting to note that 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 an 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 \( \hat{\pi}_i \) defined in (4.1) in Chapter 4.

The following is Theorem 3.2 of Newey & Smith (2004)

**Theorem 15 (Newey & Smith (2004))** Assume that (I) the parameter space \( \Theta \) is a compact set, (II) \( E [\sup_{\theta \in \Theta} \| g (\theta) \|^{\alpha}] < \infty \) 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) \( E [\sup_{\theta \in \Lambda_0} \| G (\theta) \|] < \infty \) (VII) (a) \( \text{rank} (G_0) = k \) and (b) \( G_0^T \Omega_0^{-1} G_0 \) is nonsingular. Then

\[
\begin{align*}
\frac{n^{1/2}}{2} \left[ \begin{array}{c} \hat{\lambda} \\
(\hat{\theta} - \theta_0) \end{array} \right] & \xrightarrow{d} \mathcal{N} \left( \left[ \begin{array}{cc} 0 \\
0 
\end{array} \right], \left[ \begin{array}{cc} \Psi_0 & 0 \\
0 & (G_0^T \Omega_0^{-1} G_0)^{-1} \end{array} \right] \right) \\
\frac{n^{1/2}}{2} \left( \hat{P}_p \left( \hat{\theta}, \hat{\lambda} \right) - \hat{P}_p (0) \right) & \xrightarrow{d} \chi^2_{k-1}
\end{align*}
\]

where \( \Psi_0 \) is an in (2.12).

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

\[
\begin{align*}
\sup_{\lambda \in \Lambda_n (\theta_0)} \hat{P}_p (\theta_0, \lambda) & \leq \rho_0 - \lambda' \hat{g} (\theta) + \lambda' \sum_{i=1}^n \rho_2 \left( \lambda' g_i (\theta_0) \right) g_i (\theta_0) g_i (\theta_0)' \lambda / (2n) \\
& \leq \rho_0 - \lambda' \hat{g} (\theta_0) - \lambda' C \| \lambda \|^2.
\end{align*}
\]

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

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

We next show that \( \| \hat{g} (\hat{\theta}) \| = O_p \left( n^{-1/2} \right) \) where \( \hat{\theta} = \text{arg} \min_{\theta \in \Theta} \hat{P}_p (\lambda, \theta) \) for \( \lambda \in \Lambda_n \). Let \( \bar{\lambda} = -n^\beta \hat{g} (\hat{\theta}) / \| \hat{g} (\hat{\theta}) \| \) and note that by a similar expansion as that given in (3.3)

\[
\hat{P}_p (\hat{\theta}, \bar{\lambda}) \geq \rho_0 + n^{-\beta} \| \hat{g} (\hat{\theta}) \| - C n^{-2\beta}.
\]
Combining (3.3) and (3.4) and using the definition of \( \hat{\theta} \) and the fact that \( \hat{\lambda} \) is a saddlepoint gives

\[
\rho_0 + n^{-\beta} \| \hat{g} (\hat{\theta}) \| - Cn^{-2\beta} \leq \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) \leq \sup_{\lambda \in \Lambda_n (\theta_0)} \hat{P}_\rho (\hat{\theta}, \lambda) \leq \rho_0 + O_p (n^{-1}) ,
\]

so that rearranging \( \| \hat{g} (\hat{\theta}) \| \leq O_p (n^{-\beta}) \). Let \( \tilde{\lambda} = -\varepsilon_n \hat{g} (\hat{\theta}) \) where \( \varepsilon_n \to 0 \), so that \( \tilde{\lambda} \in \Lambda_n \). Then (3.5) shows that

\[
\rho_0 - \| \tilde{\lambda} \| \| \hat{g} (\hat{\theta}) \| - C \| \tilde{\lambda} \|^2 = \rho_0 - \varepsilon_n \| \hat{g} (\hat{\theta}) \|^2 - C\varepsilon_n^2 \| \hat{g} (\hat{\theta}) \|^2 \leq \rho_0 + O_p (n^{-1})
\]

which implies that \( \| \hat{g} (\hat{\theta}) \| = O_p (n^{-1/2}) \). The consistency of \( \hat{\theta} \) follows by noting that by ULLN and T and E \( [g (\hat{\theta})] \to 0 \) and since \( E [g (\theta_0)] = 0 \) \( \hat{\theta} \) \( \hat{\theta}_0 \). The consistency of \( \hat{\lambda} \) follows by noting that \( \| \tilde{\lambda} \| \leq \| \hat{g} (\hat{\theta}) \| \xrightarrow{P} 0 \). The asymptotic distribution of \( n^{1/2} \left[ \tilde{\lambda}', (\hat{\theta} - \theta_0) \right] \) follows by mean value expansion of the FOCs \( 0 = \left[ \partial \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) / \partial \lambda , \partial \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) / \partial \theta \right] \) using ULLN, CMT noting that \( \rho_j (\tilde{\lambda}', g_i (\hat{\theta})) \xrightarrow{P} \rho_j (0) \) for \( j = 1, 2 \).

The distribution of \( 2n (\hat{P}_\rho (\hat{\theta}, \hat{\lambda}) - \bar{P}_\rho (0)) \) can be obtained by a second order Taylor expansion about 0 noting as in the proof of Theorem 4 that \( n^{1/2} \tilde{\lambda} = \Omega^{-1} n^{1/2} \hat{g} (\hat{\theta}) + o_p (1) \) and therefore

\[
2n (\hat{P}_\rho (\hat{\theta}, \hat{\lambda}) - \bar{P}_\rho (0)) = n \hat{g} (\hat{\theta})' \Omega^{-1} \hat{g} (\hat{\theta}) + o_p (1).
\]

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

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

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

The following theorem is the GEL version of Theorem 5

**Theorem 16** Under the same assumptions of Theorem 7

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

**Proof.** Almost identical to that of Theorem 5, and thus omitted. ■
CHAPTER 3. GENERALISED EMPIRICAL LIKELIHOOD AND EMPIRICAL DISCREPANCIES

3.1 Higher order asymptotic theory (II)

In this section we consider two further aspects of 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 17** Assume that (I)-(VI) of Theorem 5 hold. Furthermore assume that (VII) \(E \sup_{\beta \in N_0} \| \partial^k g(\theta)/\partial \theta^{j_1} \cdots \partial \theta^{j_k} \|^6 < \infty \) for \( k = 1, \ldots, 4 \), (VIII) for each \( \beta \in N \), \( \| \partial^4 g(\theta)/\partial \theta^{j_1} \cdots \partial \theta^{j_1} - \partial^4 g(\theta_0)/\partial \theta^{j_1} \cdots \partial \theta^{j_k} \| \leq b(z) \| \theta - \theta_0 \| \) and \( E[b(z)]^6 < \infty \), (IX) \( \rho(\cdot) \) is four times continuously differentiable with Lipschitz derivative in a neighbourhood of 0, (X) \( c \Omega = \Omega + \mathbb{P} \sum_{i=1}^{n} (z_i) = n + O_p(n^{-1}) \) where \( E[\xi(z_i)] = 0, E[||\xi(z_i)||^6] < \infty \). Then

\[
\begin{align*}
\text{Bias}(\hat{\theta}_{GMM}) &= \sum_{j=1}^{4} B_j \\
\text{Bias}(\hat{\theta}_{GEL}) &= B_1 + (1 + \rho_3/2) B_2 \\
\text{Bias}(\hat{\theta}_{EL}) &= B_1
\end{align*}
\]

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

**Proof.** Use a third order Taylor expansion of the FOCs

\[
0 = n^{1/2} \left[ \partial \hat{P}_\rho \left( \hat{\theta}, \hat{\lambda} \right) / \partial \lambda', \partial \hat{P}_\rho \left( \hat{\theta}, \hat{\lambda} \right) / \partial \theta' \right] ',
\]

invert it to obtain a stochastic expansion for \( n^{1/2} \left[ \lambda', \left( \hat{\theta} - \theta_0 \right) \right] ' \) and finally calculate expectations. \( \blacksquare \)

**Remark 18** Note that each of the \( B \) terms has an interesting interpretation: \( B_1 \) is the bias for a GMM estimator with optimal linear combination \( G' \Omega^{-1} g(\theta) \), \( B_2 \) 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 \( E[g(\theta_0) g(\theta_0)' g_j(\theta_0)] = 0 (j = 1, \ldots, l) \) then

\[
\text{Bias}(\hat{\theta}_{GEL}) = \text{Bias}(\hat{\theta}_{EL}).
\]
3.1. HIGHER ORDER ASYMPTOTIC THEORY (II)

Theorem 19 Under the same assumption of Theorem 9

$$\Xi = \Xi_{EL}$$ 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.


Remark 20 This efficiency property of EL will be shared by any estimator for which $\rho_3 = -3$ and $\rho_4 = -6$.

We now compare GEL test statistics in terms of local power. We focus as in Bravo (2003) on the ECR 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 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 in the following theorem and related proof we use the so-called index notation and indicate arrays by their elements. Thus, for any index $1 \leq r_j \leq k$ ($j = 1, 2, ..., k$), $a_r$ is an $R^k$-valued vector, $a_{rs}$ is an $R^{k \times k}$-valued matrix, etc.

Theorem 21 Assume that (I) $E [g (\theta_0)]^a < \infty$ for $a \geq 6$, (II)

$$\lim_{\|\theta\| \to \infty} \left| E \left[ \exp \left( t^{1/2} g (\theta_0) \right) \right] \right| < 1.$$

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

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

where

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

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

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

$G_{\cdot, \cdot}^- (\cdot)$ is the cumulative distribution of a noncentral chi-squared distribution with noncentrality parameter $\tau = \delta' \delta$, $\nabla^k G_{\cdot, \cdot}^- (\cdot) = \sum_{j=0}^k (-1)^{j} \binom{k}{j} G_{q+2(k-j), \cdot}^- (\cdot)$, and $c_\alpha = \Pr (\chi^2_k \geq c_\alpha) = \alpha.$
CHAPTER 3. GENERALISED EMPIRICAL LIKELIHOOD AND EMPIRICAL DISCREPANCIES

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

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

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

and cumulants $k_{r_1,...,r_v}$ given by

$$k_r = \delta_r + k_{r_1}^{11}/n^{1/2} + O \left( n^{-1} \right), \quad k_{r,s} = \delta_{rs} + k_{r,s,t}^{21}/n^{1/2} + O \left( n^{-1} \right)$$

$$k_{r,s,t} = k_{r,s,t}^{31}/n^{1/2} + O \left( n^{-1} \right), \quad k_{r_1,...,r_v} = O \left( n^{-1} \right) \quad v \geq 4,$$

where

$$k_{r_1}^{11} = \frac{1}{6} \left[ (\gamma - 1) \sum_{s=1}^{k} \alpha_{rss} + (\gamma + 2) \sum_{s,t=1}^{k} \alpha_{rst} \gamma_s \gamma_t \right],$$

$$k_{r,s}^{21} = \frac{1}{3!} \sum_{t=1}^{k} \alpha_{rst} \gamma_t /3, \quad k_{r,s,t}^{31} = \gamma \alpha_{rst}.$$

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 (3.6). 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

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

Proceeding as in the derivation of a noncentral chi-squared distribution, one obtains that the required Edgeworth expansion is given by a linear combination of terms having the form:

$$\Pr \left( CR \left( \Omega_0^{1/2} \delta/n^{1/2} \right) \geq c_{\alpha} \right) = \prod_{r=1}^{k} \left( 1 + \sum_{R_r \propto} k_{R_r} \beta_0^R \exp (w't) \right) \phi(\delta_r, 1) \, dw_r,$$
3.2 Generalised empirical likelihood for weakly identified models

where $S_{\alpha} = \{ w \in \mathbb{R}^k : w'w \geq c_{\alpha} \}$, $\delta_0^\gamma (\cdot) = \partial^\gamma (\cdot) / \partial \tau_1 ... \partial \tau_v |_{t_\gamma = 0}$, $t$ is an $\mathbb{R}^k$-valued vector of auxiliary parameters, $k_{R_v}$ are defined in (3.6), and the summation is over any index in the sets of indices $R_v = (r_1, ..., r_v)$ for $v = 1, 2, 3$. The result follows by calculating the integral and some simplifications.

Theorem 21 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 $\tilde{P}(c_\alpha, \tau) := \min P(c_\alpha, \gamma)$, where the minimum is over $\delta_\alpha = (\delta \in \mathbb{R}^k : \delta^t \delta = \tau)$.

**Theorem 22** Let $\tilde{P}_0(c_\alpha, \tau)$ denote $\tilde{P}(c_\alpha, \tau)$ for EL. Then, under the same of Theorem 11, there exists a positive $\tau_0$ such that

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

whenever $0 < \tau < \tau_0$.

**Proof.** See Bravo (2003) ■

3.2 Generalised empirical likelihood 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. 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 $\theta_\alpha = [\alpha_0', \beta_0']' \in \mathbb{R}^{p+q}$ 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)

---

2This is why in the recent econometric literature weak identification is often called “weak instruments”.
WID (1) (i) $E[\hat{g}(\theta)] = n^{-1/2}m_{1n}(\theta) + m_2(\beta)$ where $m_j(\cdot) (j = 1, 2)$ are continuous functions 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 $N_{\beta_0}$ of $\beta_0$, (iii) $\text{rank } [M_{\beta_0}] = q$ 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 non-singular for all $\theta \in A \times \beta_0$, (iii) $\sup_{\theta \in \Theta} \|\hat{\Omega}(\theta) - \Omega(\theta)\| = o_p(1)$, (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) $n^{1/2} (\hat{g}(\theta) - E[\hat{g}(\theta)]) \xrightarrow{w} W(\theta)$ 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 centered Gaussian process with covariance function $COV(W(\theta_1), W(\theta_2)) = \Delta(\theta_1, \theta_2)$, (iv) $\sup_{\theta \in A \times \beta_0} \|W(\theta)\| = O_p(1)$.

Theorem 23 (Guggenberger and Smith (2005)) Assume that (I), (III) of Theorem 15 and WID(1)-(2) hold. Then

$$\tilde{P}_p(\hat{\lambda}, \theta) \xrightarrow{d} \chi^2(c),$$

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

Proof. First note that by the results of Guggenberger & Smith (2005) $\hat{\lambda}$ 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 $\tilde{P}_p(\hat{\lambda}, \theta)$ about $\theta$ as that in Theorem 15 it follows that

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

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

$$S_p = n\hat{\lambda}' G_p(\theta) \left(G_p(\theta)' \Delta(\theta)^{-1} G_p(\theta)\right)^{-1} G_p(\theta)' \hat{\lambda},$$

and a Lagrange multiplier type statistic

$$LM_p = n\hat{g}(\theta)' \Delta(\theta)^{-1} G_p(\theta) \left(G_p(\theta)' \Delta(\theta)^{-1} G_p(\theta)\right)^{-1} G_p(\theta)' \Delta(\theta)^{-1} \hat{g}(\theta)$$

where $G_p(\theta) = \sum_{i=1}^n \theta_1 \left(\xi_i \theta_i(\theta)\right)' \partial g_i(\theta) / \partial \theta'$ . 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 $\theta_0$ $\tilde{P}_p(\hat{\lambda}, \theta_0) \xrightarrow{d} \chi^2_k$ whereas both $S_p$ and $LM_p$ converges in distribution to $\chi^2_k$. 

Chapter 4

Implied probabilities, efficient bootstrap and auxiliary information

In this chapter we consider in some detail the so-called GEL implied probabilities defined as

\[ \hat{\pi}_i = \frac{\hat{\lambda}' g_i (\hat{\theta})}{\sum_{j=1}^{n} \hat{\lambda}' g_j (\hat{\theta})}. \] (4.1)

By constructions the implied probabilities sum to one, satisfy the sample moment condition \( \sum_{i=1}^{n} \hat{\pi}_i g (\hat{\theta}) = 0 \) (w.p.a.1) when the first order conditions for \( \hat{\lambda} \) hold (by ULLN), and are positive since (cf. the proof of Theorem 15) \( \hat{\lambda}' g_i (\hat{\theta}) = o_p (1) \) uniformly in \( i \).

4.1 Asymptotic properties of GEL implied probabilities

Back & Brown (1993) showed that in overidentified moment conditions models GMM can be used to construct an estimator of the unknown distribution \( F \) of the observations that is asymptotically efficient compared to its nonparametric maximum likelihood estimator -the empirical distribution function . In this section we show that the GEL IP-based estimator for \( F \) is also asymptotically
efficient. Let
\[ \hat{F}_n (z) = \sum_{i=1}^{n} \pi_i I \{ z_i \leq z \} \]  \hspace{1cm} (4.2)

**Theorem 25** Under the same assumptions of Theorem 15,
\[ n^{1/2} \left( \hat{F}_n (z) - F (z) \right) \overset{d}{\to} N \left( 0, F (z) \left( 1 - F (z) \right) \right. \]  \hspace{1cm} B (z)' \Psi_0 B (z), \]  \hspace{1cm} (4.3)

where \( B (z) = Eg_i (z_i, \theta_0) I \{ z_i \leq z \} \) and \( \Psi_0 \) is as (2.12).

**Proof.** Since \( \hat{\lambda} g_i (\hat{\theta}) = o_p (1) \) uniformly in \( i \), by a mean value expansion
\[ \rho_1 \left( \hat{\lambda} g_i (\hat{\theta}) \right) = -1 + \rho_2 \left( X g_i (\hat{\theta}) \right) \hat{\lambda} g_i (\hat{\theta}) = -1 - \hat{\lambda} g_i (\hat{\theta}) + o_p (1). \]
Similarly
\[ \frac{1}{n} \left( \hat{\lambda} g_i (\hat{\theta}) \right) = -1/n \left( 1 + \frac{1}{n} \rho_2 \left( X g_i (\hat{\theta}) \right) g_i (\hat{\theta})' / n \right) \hat{\lambda} = -1/n (1 + O_p (n^{-1})), \]
so that
\[ \max_i \left| \hat{\pi}_i - 1/n \left( 1 + \hat{\lambda} g_i (\hat{\theta}) \right) \right| = O_p (n^{-1}), \]  \hspace{1cm} (4.4)
and therefore
\[ n^{1/2} \left( \hat{F}_n (z) - F (z) \right) \overset{a}{=} n^{1/2} \left( \hat{F}_n (z) - F (z) \right) - n^{1/2} \sum_{i=1}^{n} I \{ z_i \leq z \} \hat{\lambda} g_i (\hat{\theta}) / n \]
\[ \overset{(4.5)}{=} \sum_{i=1}^{n} I \{ z_i \leq z \} g_i (\hat{\theta}) / n \overset{p}{\to} B (z), \]
so that by CLT and CMT the result follows since
\[ V \left[ n^{1/2} \left( F_n (z) - F (z) - \hat{\lambda} B (z) \right) \right] = nV ((F_n (z) - F (z))) + nV \left( \hat{\lambda} B (z) \right) - 2E \left( n (F_n (z) - F (z)) \hat{\lambda} B (z) \right) = nV ((F_n (z) - F (z))) - nV \left( \hat{\lambda} B (z) \right). \]

To see why \( \hat{F}_n (z) \) is asymptotically efficient with respect to the empirical distribution function (EDF) - which is the typical nonparametric estimator used to estimate \( F (z) \) note that
\[ n^{1/2} \left( \hat{F}_n (z) - F (z) \right) \overset{d}{\to} N \left( 0, F (z) \left( 1 - F (z) \right) \right) \]  \hspace{1cm} (4.6)
The reason for the efficiency of \( \hat{F}_n (z) \) is that it incorporates the information contained in the moment condition model, that is that the model is overidentified. Indeed \( \hat{F}_n \) achieves the semiparametric efficiency bound of Brown & Newey (1998) for estimators of \( F \) under the moment condition (2.7).
4.1. ASYMPTOTIC PROPERTIES OF GEL IMPLIED PROBABILITIES

Theorem 25 shows the pointwise asymptotic normality of \( \hat{F}_n(z) \). The following theorem establishes the weak convergence in the Skorohod space \( D[-\infty, \infty] \) of the GEL IP modified empirical process

\[
n^{1/2} \left( \hat{F}_n - F \right) = G_n.
\]

**Theorem 26** Under the same assumptions of Theorem 15, then

\[
G_n \xrightarrow{w} W
\]

where \( W \) is a centered Gaussian process with continuous sample paths and covariance function

\[
\text{COV}(W(x), W(y)) = F(x \wedge y) - F(x) F(y) - B(z)' \Psi_0 B(z)
\]

\[
E \left[ g_i(\theta_0) I(z_i \leq x) \right]' \Psi_0 E \left[ g_i(\theta_0) I(z_i \leq y) \right]
\]

where \( B(z)' \Psi_0 B(z) \) is defined in Theorem 25.

**Proof.** By Lemma 28 in the Appendix

\[
G_n(z) = \sum_{i=1}^n X_i(z) / n^{1/2} + o_p(1) = G_n + Q_n(z)
\]

where \( Q_n(z) = \sum_{i=1}^n E \left[ g_i(\theta_0) I \{ z_i \leq z \} \right] \Psi_0 g_i(\theta_0) / n^{1/2} \). It is well-known that \( G_n \) is tight in \( D[-\infty, \infty] \), and therefore for every \( \varepsilon > 0 \) there exist a finite \( M_\varepsilon \) such that

\[
\Pr \left( \sup_{z \in \mathbb{R}} |G_n(z)| > M_\varepsilon/2 \right) \leq \varepsilon/2,
\]

and

\[
\lim_{n \to \infty} \limsup_{\delta \to 0} \Pr \left( \sup_{|z-y| < \delta} \left| G_n(x) - G_n(y) \right| > \varepsilon/2 \right) = 0.
\]

For the same \( M_\varepsilon \), Chebychev’s inequality gives

\[
P \left( \sup_{z \in \mathbb{R}} |Q_n(z)| > M_\varepsilon/2 \right) \leq (4I^2/M_\varepsilon^2) \sum_{j=1}^l \sum_{k=1}^l |\Psi_{0jk}|^2 \left( E \left[ g_j(\theta_0) \right] \right)^2 E g_k(\theta_0)^2 \leq \varepsilon/2
\]

where \( \Psi_{0jk} \) denotes the \( jk \)th element of the matrix \( \Psi_0 \). Moreover

\[
\Pr \left( \sup_{|x-y| < \delta} \left| Q_n(x) - Q_n(y) \right| > \varepsilon/2 \right) \leq \sup_{|x-y| < \delta} \left| F(x) - F(y) \right| \left( 4I^2/\varepsilon^2 \right) \sum_{j=1}^l \sum_{k=1}^l |\Psi_{0jk}|^2 \left( E \left[ g_j(\theta_0) \right] \right)^2 E g_k(\theta_0)^2.
\]
Since $F$ is continuous

$$\lim_{\delta \to 0} \lim_{n \to \infty} \Pr \left( \sup_{|x-y|<\delta} |Q_n(x) - Q_n(y)| > \varepsilon / 2 \right) = 0.$$ 

Thus combining the above results

$$\lim_{\delta \to 0} \lim_{n \to \infty} \Pr \left( \sup_{|x-y|<\delta} \left| \sum_{i=1}^n X_i(x) / n^{1/2} - \sum_{i=1}^n X_i(y) / n^{1/2} \right| > \varepsilon \right) = 0,$$

that is the process $\sum_{i=1}^n X_i(\cdot) / n^{1/2}$ is tight in $D [-\infty, \infty]$. By CLT and CMT

$$\sum_{i=1}^n X_i(z) / n^{1/2} \overset{d}{\to} N (0, F(z) (1 - F(z)) - \Lambda_0),$$

where $\Lambda_0 = E [g_i(\theta_0) I \{z_i \leq z\}] \Psi_0 E [g_i(\theta_0) I \{z_i \leq z\}]$, which proves the result.

4.1.1 Appendix

**Lemma 27** Let $h(\cdot)$ be an $\mathbb{R}$-valued function such that $E|h(z)| = Q < \infty$, and let $H_n(x) = \sum_{i=1}^n h(z_i) I \{z_i \leq x\} / n$ and $H(x) = E[h(z) I \{z \leq x\}]$. Then

$$\sup_{x \in \mathbb{R}} |H_n(x) - H(x)| = o_p(1).$$

**Proof.** Similar to Glivenko-Cantelli theorem and thus omitted.

**Lemma 28** Under the same assumptions of Theorem 15, then

$$\hat{F}_n(z) - F(z) = \hat{X}(z) + \xi_n(z)$$

where

$$X_i(z) = I \{z_i \leq z\} - F(z) - E [g_i(\theta_0) I \{z_i \leq z\}] \Psi_0 g_i(\theta_0),$$

and $\sup_{z \in \mathbb{R}} |\xi_n(z)| = o_p(n^{-1/2}).$

**Proof.** As in Theorem 25

$$\hat{F}_n(z) - F(z) = \hat{X}(z) + \sum_{j=1}^4 \xi_{jn}(z)$$
4.2. EFFICIENT BOOTSTRAP IN MOMENT CONDITIONS MODELS

where

\[
\begin{align*}
\xi_{1n}(z) &= -E \left[ g_i(\theta_0)' I \{ z_i \leq z \} \right] \left( \hat{\lambda} - \Psi_0 \hat{g}(\theta_0) \right), \\
\xi_{2n}(z) &= -\hat{\lambda}' (\hat{g}(\theta_0) - E [g(\theta_0)]) I \{ z_i \leq z \}, \\
\xi_{3n}(z) &= -\sum_{i=1}^{n} \left[ \hat{\lambda}' g_i(\hat{\theta}) \right]^2 I \{ z_i \leq z \} / n, \\
\xi_{4n}(z) &= -\hat{\lambda}' (\hat{g} - \hat{g}(\theta_0)) I \{ z_i \leq z \}.
\end{align*}
\]

Note that \( \| \hat{\lambda} - \Psi_0 \hat{g}(\beta_0) \| = o_p(n^{-1/2}) \) and therefore \( \sup_{z \in \mathbb{R}} | \xi_{1n} | = o_p(n^{-1/2}) \).
Furthermore by Lemma 27

\[
\sup_{z \in \mathbb{R}} | (\hat{g}_j(\theta_0) - E [g_j(\theta_0)]) I \{ z_i \leq z \} | = o_p(1) \ (j = 1, ..., l)
\]

and therefore \( \sup_{z \in \mathbb{R}} | \xi_{2n} | = o_p(n^{-1/2}) \). In addition by ULLN \( \sum_{i=1}^{n} \| g_i(\hat{\theta}) \|^2 / n = O_p(1) \) and thus \( | \xi_{3n} | = O_p(n^{-1}) \). Finally by Lemma 2.4 of Newey & McFadden (1994), mean value theorem, CMT and

\[
\| \hat{g}(\hat{\theta}) - \hat{g}(\theta_0) \| \leq \| \hat{G}(\hat{\theta}) \| \| \hat{\theta} - \theta_0 \| = O \left( n^{-1/2} \right)
\]

and thus \( | \xi_{4n} | = O_p(n^{-1}) \).

4.2 Efficient bootstrap in moment conditions models

In this section we show how the GEL IP can be used to bootstrap GMM estimators. The resulting bootstrap procedure, called efficient by Brown & Newey (2002), delivers the same type of asymptotic refinements delivered by the standard bootstrap, but has the advantage of not requiring the recentring of the resampled moment conditions.

Remark 29 The efficient bootstrap is an example of the so-called “biased” or b-bootstrap (Hall & Presnell 1999). In the b-bootstrap the resampling probabilities are chosen as to minimise the distance of the weighted bootstrap distribution from the usual uniform distribution subject to a given constraint that is designed to improve statistical performance, such as bias reduction or variance stabilising.

To describe the efficient bootstrap algorithm let \( \hat{\beta} \) denote an efficient GMM estimator such as the two step GMM or any GEL estimators, and let \( S(\theta) \) denote a given statistic of interest.
Algorithm 30 (Efficient Bootstrap) (1) Compute the IP $\hat{\pi}_i$. (2) Draw with replacement $n$ i.i.d. observations $z_i^*$ from the distribution with $\Pr(z = z_i) = \hat{\pi}_i$. (3) Compute a statistic of interest $S^*(\theta^*)$. (4) Repeat step (3) $m$ times to obtain the bootstrap distribution of $S^*(\theta^*)$.

The bootstrap estimator of the distribution of $S$ can be used to form critical values for test statistics and construct confidence regions (intervals) in the usual way. The following theorem establishes the consistency of the efficient bootstrap GMM estimator and $J$ statistic for overidentified restrictions. Let

$$\hat{\theta}^* = \arg \min_{\theta \in \Theta} \hat{\theta}^* \hat{W}^* \hat{\theta}^*$$

where $\hat{W}^*$ is a possibly (bootstrapped) random symmetric matrix assumed positive semi-definite (w.p.a 1), denote the efficient bootstrap estimator and let

$$J_n^* = n\hat{g}^*(\hat{\theta}^*)^{-1} \hat{\theta}^*$$

denote the bootstrapped $J$ statistic. Let $J_W(\theta) = g(\theta)' \hat{W} g(\theta)$ where $\hat{W}$ is a possibly random positive semidefinite $l \times l$ matrix.

Theorem 31 Assume that (I) the parameter space $\Theta$ is a compact set, (II) for all $\zeta > 0 \inf_{\|\theta - \theta_0\| > \zeta} \|E[J_W(\theta)]\| \geq \varepsilon(\zeta) > 0$, (III) $\sup_{\theta \in \Theta} \|\hat{J}_W(\theta) - E[J_W(\theta)]\| = o_p(1)$, (IV) $\theta_0 \in \text{int} \{\Theta\}$, (V) $\sup_{\theta \in \mathcal{N}_0} \|\hat{G}(\theta) - E[\hat{G}(\theta)]\| = o_p(1)$, (VI) $E(G_0)$ is nonsingular where $G_0 = G(\theta_0)$, (VII) $n^{1/2} \hat{g}(\hat{\theta}_0) \overset{d}{\rightarrow} N(0, \Omega_0)$ where $\Omega_0 = E[\hat{g}(\theta_0)g(\theta_0)']$. Then

$$\sup_{x \in \mathbb{R}^k} \Pr^*(n^{1/2}(\hat{\theta}^* - \hat{\theta}) < x) - \Pr(n^{1/2}(\hat{\theta} - \theta_0) < x) = o_p(1)$$

$$\sup_{x \in [x_0, \infty)} |\Pr^*(J^* < x) - \Pr(J < x)| = o_p(1)$$

Proof. We first show that $\hat{\theta}^* - \hat{\theta} = o_p^*(1)$ in probability. This follows by Lemma 35 and the identification condition (II). Next note that

$$n^{1/2} \left( \hat{g}^*(\hat{\theta}) - \hat{g}(\hat{\theta}) \right) = n^{1/2} (\hat{g}^*(\theta_0) - E^* g_i(\theta_0)) -$$

$$n^{1/2} \left( \hat{g}(\hat{\theta}) - \hat{g}(\theta_0) \right) + n^{1/2} (\hat{g}^*(\hat{\theta}) - \hat{g}^*(\theta_0)) + o_p^*(1) = \sum_{j=1}^3 A_j.$$
4.2. EFFICIENT BOOTSTRAP IN MOMENT CONDITIONS MODELS

By the bootstrap CLT (Bickel & Freedman 1981) \( A_1 \overset{d}{\rightarrow} N(0, \Omega_0) \) in probability, while by mean value expansion

\[
\|A_2 + A_3\| \leq \sup_{\theta \in \Theta_0} \left\| \hat{G}^* (\theta) - \hat{G} (\theta) \right\| n^{1/2} \left( \hat{\theta} - \theta_0 \right) = o_p (1) \]

Finally by mean value expansion, CMT and Lemma 33 of the bootstrap FOCs

\[
a_p^* (1) = \hat{G}^* (\hat{\theta}^*) \tilde{W} \hat{g}^* (\hat{\theta}^*) \quad \text{in probability}
= G_0 W \left[ n^{1/2} \hat{g}^* (\hat{\theta}) + G_0 n^{1/2} (\hat{\theta}^* - \hat{\theta}) \right] \quad \text{in probability}
\]

which shows that

\[
n^{1/2} (\hat{\theta}^* - \hat{\theta}) \overset{d}{\rightarrow} N \left( 0, (G_0 W G_0)^{-1} G_0 W \Omega_0 W G_0 (G_0 W G_0)^{-1} \right) \]

which implies the first conclusion. The second conclusion follows by similar arguments.

Theorem 31 shows the asymptotic validity of the efficient bootstrap, but does not establish any asymptotic refinements. As argued by Brown & Newey (2002) the asymptotic refinements of the efficient bootstrap and the efficiency of efficient bootstrap over the traditional bootstrap can be explained in terms of Edgeworth expansions. Using a similar notation as that used by Brown & Newey (2002) let

\[ H_n \left( \hat{\theta}, F \right) = \Pr \left( S (\hat{\theta}) \leq c | F \right) \]

denote the distribution of \( S (\theta) \) and note that the efficient bootstrap estimator is \( H_n \left( \hat{\theta}^*, \hat{F}_n \right) \). By a formal Edgeworth expansion of \( H_n \left( \hat{\theta}, F \right) \)

\[ H_n \left( \hat{\theta}, F \right) = H_\infty + n^{-\alpha} R_1 (\theta_0, F) + n^{-\beta} R_2 (\theta_0, F) + o_n \left( n^{-\beta} \right) \]

where \( H_\infty \) is the asymptotic distribution of \( S (\theta_0) \) (assumed to exist and be independent of \( \theta_0 \) and \( F \)), and \( \beta \geq \alpha + 1/2 \). Using the same expansion for \( \Pr^* \left( S (\hat{\theta}^*) \leq c | \hat{F}_n \right) \) and subtracting gives the efficient bootstrap error

\[ H_n \left( \hat{\theta}^*, \hat{F}_n \right) - H_n \left( \hat{\theta}, F \right) = n^{-\alpha} \left( R_1 \left( \hat{\theta}, \hat{F}_n \right) - R_1 (\theta_0, F) \right) + o_p \left( n^{-\alpha} \right) = o_p \left( n^{-\alpha} \right) . \]  
(4.6)

Since \( H_\infty - H_n \left( \hat{\theta}, F \right) = O \left( n^{-1/2} \right) \) (4.6) shows that the efficient bootstrap provides the same type of asymptotic refinement provided by the standard bootstrap. A refinement of this approach can be used to show that the efficient bootstrap is efficient in a certain sense. Assume that \( \left| R_2 \left( \hat{\theta}, \hat{F} \right) - R_2 (\theta_0, F) \right| = o_p (1) \), and note that

\[ n^{\alpha + 1/2} \left( H_n \left( \hat{\theta}^*, \hat{F}_n \right) - H_n \left( \hat{\theta}, F \right) \right) = n^{1/2} \left( R_1 \left( \hat{\theta}, \hat{F} \right) - R_1 (\theta_0, F) \right) + o_p (1) , \]
and that \( n^{1/2} \left( R_1 \left( \hat{\theta}, \hat{F}_n \right) - R_1 \left( \theta_0, F \right) \right) \) is asymptotically normal with a variance that is directly related to the asymptotic variance of \( R_1 \left( \hat{\theta}, \hat{F}_n \right) \). The more efficient \( R_1 \left( \hat{\theta}, \hat{F}_n \right) \) the closer will be the bootstrap remainder to zero in large samples. This is Brown & Newey (2002) notion of efficiency of bootstrap, and it shows that any GEL based bootstrap is efficient relative to the standard bootstrap because its approximation has a smaller variance.

**Remark 32** It should be noted that the efficiency comparisons do not necessarily carry over to asymptotic refinements for test statistics (i.e. the fact that we can estimate more precisely the asymptotic distribution of a statistic does not imply that we are able to improve upon the order the approximation error).

### 4.2.1 Appendix

**Lemma 33** Under the same assumptions of Theorem 16, then

\[
E^* \sup_{\theta \in \Theta} \| g^* (\theta) \| < \infty \quad \text{in probability.}
\]

**Proof.** \( E^* \sup_{\theta \in \Theta} \| g^* (\theta) \| = \sum_{i=1}^n \hat{\pi}_i \sup_{\theta \in \Theta} \| g_i (\theta) \| \xrightarrow{p} E \sup_{\theta \in \Theta} \| g (\theta) \| < \infty. \)

**Lemma 34** Under the same assumptions of Theorem 31, then for each \( \theta \)

\[
\| \hat{g}^* (\theta) - E^* [ g^* (\theta) ] \| = o_p (1) \quad \text{in probability.}
\]

**Proof.** By T and

\[
\| \hat{g}^* (\theta) - E^* [ g^* (\theta) ] \| \leq \| \hat{g}^* (\theta) - \hat{g} (\theta) \| + \max_i | \lambda_i | g_i (\theta) \| \hat{g} (\theta) \|
\]

\[
= o_p (1) \quad \text{in probability} + o_p (1)
\]

where the last line follows by M. \( \blacksquare \)

**Lemma 35** Under the same assumptions of Theorem 31, then

\[
\sup_{\theta \in \Theta} \| \hat{g}^* (\theta) - E^* [ g^* (\theta) ] \| = o_p (1) \quad \text{in probability.}
\]

**Proof.** Note that by assumption (I), Lemma 34 and the continuity of \( g (\cdot) \) it is enough to show that \( \hat{g}^*_n (\theta) = \hat{g}^* (\theta) - E^* [ g^* (\theta) ] \) is stochastically equicontinuous. To this end let

\[
g^*_n (\theta) = \sup_{\theta \in \Theta} \sup_{\theta' \in N(\theta, \delta)} \sum_{i=1}^n \| g_i (\theta') - g_i (\theta) \|
\]
and note that by Lemma 33 $E^* (g_{n\delta}^*) \leq 2E \sup \|g (z_i, \theta)\| < \infty$ and therefore $E^* (g_{n\delta}^*) \xrightarrow{p} 0$ by dominated convergence. It then follows that
\[
\sup_{\theta \in \Theta} \sup_{\theta' \in \mathcal{N}(\theta, \delta)} \left\| \bar{g}_n^* (\theta') - \bar{g}_n^* (\theta) \right\| \leq g_{n\delta}^* + E^* (g_{n\delta}^*)
\]
and thus by M
\[
\lim_{n \to \infty} \Pr \left( \sup_{\theta \in \Theta} \sup_{\theta' \in \mathcal{N}(\theta, \delta)} \left\| \bar{g}_n^* (\theta') - \bar{g}_n^* (\theta) \right\| > \varepsilon \right) \leq \lim_{n \to \infty} E^* (g_{n\delta}^*) / \varepsilon \xrightarrow{p} 0.
\]

4.3  Efficient Z-estimation with auxiliary information

In many situations of practical interest we may have some auxiliary information about the otherwise unknown distribution $F$ of the sample. For example we might know the probability that the observed data belong to a certain part of the sample space, or that $F$ has given known moments (joint or marginal), or that is symmetric around a certain constant. This information is often available from auxiliary data such as national statistics or the census. In these situations we might expect that incorporating such information into the estimation process can reduce the bias and increase the efficiency of the parameter estimates (see for example Imbens & Lancaster (1994) and Hellerstein & Imbens (1999)).

In this section we show how GEL IP can be used to incorporate auxiliary information into a Z-estimation process. We consider approximate Z-estimator defined as
\[
\left\| \hat{g} (\tilde{\theta}) \right\| \leq \inf_{\theta \in \Theta} \left\| \bar{g} (\theta) \right\| + o_p \left( n^{-1/2} \right),
\]
and assume that there exists auxiliary information about $F$ available in the moment condition form
\[
E [h (z)] = 0
\]
where $h (\cdot) \in \mathbb{R}^p$. Then using a set of GEL IP consistent with (4.8), we can define a new class of GEL-based Z-estimators $\hat{\theta}_g$ as
\[
\left\| \hat{g}_g (\tilde{\theta}) \right\| \leq \inf_{\theta \in \Theta} \left\| \bar{g}_g (\theta) \right\| + o_p \left( n^{-1/2} \right).
\]
where $\hat{g}_g (\theta) = \sum_{i=1}^n \hat{\pi}_i g_i (\theta)$. The following theorem establishes consistency and asymptotic normality of $\hat{\theta}_g$; let $E [h (z) h (z)'] = \Sigma.$
Theorem 36 Under assumptions (I)-(III) of Theorem 31, $\hat{\theta}_z \xrightarrow{p} \theta_0$. Assume further that (IV) $\theta_0 \in \text{int} \{ \Theta \}$, (V) there exists a finite nonsingular matrix $\Gamma_0$ such that $\lim_{n \to 0} [\theta_0 - \theta_0] h(\theta) \rightarrow 0$, (VI) for all positive $\delta_n \to 0 \sup_{\| \theta - \theta_0 \| \leq \delta_n} \| \hat{g}(\theta) - g(\theta) - \hat{\Gamma}(\theta_0) \| = o_p(n^{-1/2})$ (VII) there exists a neighbourhood $N_0$ of $\theta_0$ such that $E \sup_{N_0} \| g(\theta) h(z) \| < \infty$ (VIII) $n^{1/2}\hat{g}(\theta_0) \rightarrow N(0, \Omega_0)$. Then under the assumptions of Theorem 1

$$n^{1/2} \left( \hat{\theta}_z - \theta_0 \right) \rightarrow N(0, \Xi_0)$$

where

$$\Xi_0 = \Gamma_0^{-1} \left\{ \Omega_0 - E \left[ g(\theta_0) h(z) \right] \Sigma_0^{-1} E \left[ g(\theta_0) h(z) \right] \right\} \left( \Gamma_0 \right)^{-1}.$$  

Proof. Note that

$$\sup_{\theta \in \Theta} \| \hat{g}(\theta) \| \leq \sup_{\theta \in \Theta} \| g(\theta) \| (1 + o_p(1))$$

and the rest of the proof is as in Theorem 31. To establish the asymptotic normality let $dG_n(z) = d\hat{F}_n(z) - dF(z)$, and $h(z_i) = h_i$. Note that

$$\hat{g}_z(\theta) = \Gamma(\theta - \theta_0) + o(\| \theta - \theta_0 \| + g(\theta) \left( 1 + \hat{\lambda}' h_i \right) + o_p(1) \right) = H_z(\theta) + o(\| \theta - \theta_0 \|) +$$

$$\int (g(\theta) - g(\theta_0)) dG_n(z) + \hat{\lambda} \sum [g_i(\theta) - g_i(\theta_0)] h_i/n + o_p(1)$$

where $H_z(\theta) = \Gamma(\theta - \theta_0) + \hat{g}(\theta_0) \left( 1 + \hat{\lambda}' h_i \right)$, and note that by (4.9) (V), (VI) and (VII) $n^{1/2} \left( \hat{\theta}_z - \theta_0 \right) = O_p(1)$. Then

$$\left\| n^{1/2} \left( \hat{g}_z(\hat{\theta}) - H_z(\hat{\theta}) \right) \right\| \leq o_p(1) +$$

$$\sup_{\| \theta - \theta_0 \| \leq \delta_n} \left\| n^{1/2} \left[ \int (g_i(\theta) - g_i(\theta_0)) dG_n(z) \right] \right\| +$$

$$\left\| n^{1/2} \hat{\lambda} \right\| \left\| \sum [g_i(\hat{\theta}_z) - g_i(\theta_0)] h_i/n \right\| = A_1 + A_2.$$

(4.10)

By (VI) $A_1 = o_p(1)$ while by continuity of $g(\theta)$ and the consistency of $\hat{\theta}$ there exists a $\delta_n \to 0$ such that $\sup_{\| \theta - \theta_0 \| \leq \delta_n} \| (g_i(\theta) - g_i(\theta_0)) h_i \| = o_p(1)$. Then by (VI) and dominated convergence $E \sup_{\| \theta - \theta_0 \| \leq \delta_n} \| (g(\theta) - g(\theta_0)) h \| \to 0$ so that by $T$ and $M$

$$\left\| \sum [g_i(\hat{\theta}_z) - g_i(\theta_0)] h_i/n \right\| \leq \sum \sup_{\| \theta - \theta_0 \| \leq \delta_n} \| (g_i(\theta) - g_i(\theta_0)) h_i \| / n = o_p(1),$$

and $A_2 = o_p(1)$. Thus $n^{1/2} \hat{g}_z(\hat{\theta})$ is asymptotically equivalent to $n^{1/2} H_z(\hat{\theta})$. Let $\hat{\theta} := \arg \min_{\theta} \| H_z(\theta) \|$ and note that $\left\| n^{1/2} H_z(\hat{\theta}) \right\| = \left\| n^{1/2} H_z(\hat{\theta}) \right\| +$
4.3. EFFICIENT Z-ESTIMATION WITH AUXILIARY INFORMATION

\[ o_p(1), \text{which implies that} \]

\[ \left\| n^{1/2} \Gamma \left( \hat{\theta} - \bar{\theta} \right) \right\| \leq \| \Gamma \| \left\| n^{1/2} \left( \hat{\theta} - \bar{\theta} \right) \right\| = C o_p(1) \]

and hence \( \left\| n^{1/2} \left( \hat{\theta} - \bar{\theta} \right) \right\| = o_p(1) \). Thus the distribution of \( \hat{\theta} \) is asymptotically equivalent to that of \( \bar{\theta} \). □

Example 37 Let \( z = [y, x']' \) and \( q_p(y|x) := \inf \{ y : F(y|x) \geq p \} = x' \theta_{p0} \) denote the \( p \)th (\( 0 < p < 1 \)) quantile of \( y \) conditional on \( x \). The GEL based quantile regression estimator \( \hat{\theta}_{p \hat{\pi}} \) for \( \theta_p \) solves \( \hat{g}_\pi \left( \hat{\theta}_{p \hat{\pi}} \right) = 0 \), where \( g_i(\theta_p) := \hat{g}_i x_i \text{sign}_p \{ y_i - x_i' \theta_p \} \), and \( \text{sign}_p \{ \cdot \} = p I \{ \cdot \leq 0 \} - (1 - p) I \{ \cdot \geq 0 \} \). Let \( \varepsilon = y - x' \theta_{p0} \). Under standard regularity conditions it is possible to show that the GEL based quantile regression estimator \( \hat{\theta}_{p \hat{\pi}} \)

\[ n^{1/2} \left( \hat{\theta}_{p \hat{\pi}} - \theta_{p0} \right) \xrightarrow{d} N(0, \Xi_0), \]

where

\[ \Xi_0 = \Gamma^{-1} \left\{ p (1 - p) E(xx') - E \left[ \text{sign}_p \{ \varepsilon \} xg(z)' \right] \Sigma^{-1} E \left[ \text{sign}_p \{ \varepsilon \} xg(z)' \right]' \right\} \Gamma^{-1}, \]

\[ \Gamma = -E \left[ f_\varepsilon(0|x) xxx' \right]. \] The following table reports the finite sample bias, variances and efficiency assuming \( E[g(z)] = [E(y) - \mu_{10}, E(y^2) - \mu_{20}]' = 0 \). We set \( \theta_0 = [1, 0.5]' \), \( x_i = [1, x_{1i}]' \), and \( x_{1i} \) is \( N(0, 1) \).
Finite sample bias $B$, variances $V$, and efficiency $E$ of $\hat{\theta}_p$, $\hat{\theta}_{1p}^{GMM}$, and $\hat{\theta}_{1p}^{GEL}$ in quantile regression model with for $N(0, 1)$ observations and

<table>
<thead>
<tr>
<th>$p$</th>
<th>$N(0, 1)$ errors</th>
<th>$t(4)$ errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td>$b_{1p}$</td>
<td>$\hat{\theta}_{1p}^{GMM}$</td>
</tr>
<tr>
<td>$\hat{b}_{1p}$</td>
<td>$\hat{b}_{2p}$</td>
<td>$\hat{b}_{2p}$</td>
</tr>
<tr>
<td>$b_{1p}$</td>
<td>.0079 .0223</td>
<td>.0373 .0029</td>
</tr>
<tr>
<td>$E$</td>
<td>1 1</td>
<td>1.079 1.097</td>
</tr>
<tr>
<td>.50</td>
<td>$b_{1p}$</td>
<td>$\hat{b}_{2p}$</td>
</tr>
<tr>
<td>$b_{1p}$</td>
<td>.0020 .0008</td>
<td>-.0004 .0036</td>
</tr>
<tr>
<td>$E$</td>
<td>1 1</td>
<td>1.309 1.188</td>
</tr>
<tr>
<td>.75</td>
<td>$b_{1p}$</td>
<td>$\hat{b}_{2p}$</td>
</tr>
<tr>
<td>$b_{1p}$</td>
<td>.0063 .0010</td>
<td>-.0062 .0028</td>
</tr>
<tr>
<td>$E$</td>
<td>1 1</td>
<td>1.137 1.104</td>
</tr>
<tr>
<td>.25</td>
<td>$b_{1p}$</td>
<td>$\hat{b}_{2p}$</td>
</tr>
<tr>
<td>$b_{1p}$</td>
<td>.0089 .0086</td>
<td>-.0069 .0036</td>
</tr>
<tr>
<td>$E$</td>
<td>1 1</td>
<td>1.064 1.025</td>
</tr>
<tr>
<td>.50</td>
<td>$b_{1p}$</td>
<td>$\hat{b}_{2p}$</td>
</tr>
<tr>
<td>$b_{1p}$</td>
<td>.0124 .0032</td>
<td>.0012 .0028</td>
</tr>
<tr>
<td>$E$</td>
<td>1 1</td>
<td>1.063 1.089</td>
</tr>
<tr>
<td>.75</td>
<td>$b_{1p}$</td>
<td>$\hat{b}_{2p}$</td>
</tr>
<tr>
<td>$b_{1p}$</td>
<td>.0392 .0405</td>
<td>.0348 .0376</td>
</tr>
</tbody>
</table>

GMM Efficient GMM, EL Empirical likelihood, EU Euclidean likelihood, NT Nonparametric tilting
Chapter 5

Generalised empirical likelihood and smoothing

This chapter adapts GEL to curve estimation problems such as density estimation and nonparametric regression. Let $K_h(z) := K(z/h)$ denote a kernel function and $h = h(n)$ the bandwidth parameter.

5.1 Density estimation

Let $(z_i)_{i=1}^n$ be i.i.d. observations from an unknown distribution $F$ with density $f$. The kernel density estimator for $f$ at an arbitrary fixed point $z$ is

$$\hat{f}(z) = \frac{1}{nh} \sum_{i=1}^n K_h(z_i - z).$$

Often confidence intervals for $f(z)$ are required (for example to test a hypothesis about $f(z)$). One possible way to construct such intervals is to use the bootstrap. However because of the bias in the kernel estimate the smoothing bandwidth $h$ should be of order $n^{-1/(r+1)}$ (where $r$ is the order of the kernel) to obtain valid confidence intervals with a two-sided percentile $t$ method. This results typically in poor empirical coverage, because the required undersmoothing affects greatly the precision of the variance estimation needed for the percentile $t$ method. To overcome this difficulty Chen (1996) propose to use EL to construct confidence intervals for $f(z)$. The rationale behind this proposal is that EL does not require explicit variance estimation and therefore this should result is better confidence intervals.
Let \( g_{ih}(f(z)) = K_{ih}(z_i - z)/h - f(z) \), and note that assuming \( f \) to be sufficiently smooth

\[
E[K_{ih}(z_i - z)/h] = f(z) + kd^r f(z)/dz^r h^r/r! + o(h^r).
\]

Thus the “moment indicator” \( g_{ih}() \) is biased, and therefore confidence intervals do not have the right coverage. One way to solve the problem is to correct explicitly the confidence interval by shifting it by \( kd^r f(z)/dz^r h^r/r! \). Another way is to undersmooth. The following theorem show that GEL with undersmoothing can be used to obtain confidence intervals for \( f(z) \) with asymptotic correct coverage. The theorem generalises some of Chen’s (1996) results. Let

\[
I_\alpha = \{ f(z) \mid \left( \hat{P}_\rho \left( f(z), \hat{\lambda} \right) - \hat{P}_\rho (0) \right) \leq c_\alpha \}
\]

where \( Pr(\chi_1^2 \leq c_\alpha) = \alpha \) denote a GEL confidence interval with nominal coverage \( \alpha \).

**Theorem 38** Assume that (I) the \( r \)th \((r \geq 2)\) order kernel \( K(\cdot) \) is bounded and has a compact support, (II) \( f \) has continuous derivatives up to the \( r \)th order in a neighbourhood of \( z \) and \( f(z) > 0 \), (III) \( h \to 0 \), \( nh \to \infty \), and \( n^{1/2} h^{r+1/2} \to 0 \). Then

\[
Pr(f(z) \in I_\alpha) = \alpha + o(1).
\]

**Proof.** Let \( g_{ih}(f(z)) := g_{ih}, \mu_j = E\left(g_{ih}^j\right) (j = 1, 2, \ldots) \) and let \( \delta_n = n^{-1/2} \max_i |g_{ih}| \). Define \( \Lambda_n = \{ |\lambda| \mid |\lambda| \leq n^{-1/2} \delta_n^{-1/2} \} \) and note that \( \max_i |g_{ih}| = o_p (n^{1/2}) \) provided \( n^{1/2} h \to \infty \). Then \( \max_i \sup_{\lambda \in \Lambda_n} |\lambda g_{ih}| \leq \delta_n^{1/2} = o_p (1) \) and by the usual arguments

\[
0 = \hat{P}_\rho \left( f(z), \hat{\lambda} \right) \leq 2\left| \hat{\lambda} - \hat{\lambda}^2 \hat{g}_h \right|.
\]

Note that by LLN \( \hat{\lambda}^2 \hat{g}_h \rho_{\mu_2} = O(h^{-1}) \) and that \( \hat{g}_h = O_p\left(n^{-1/2} + h^r\right) \). Therefore \( |\hat{\lambda}| \leq |\hat{g}_h|/\mu_2 = O_p\left(n^{-1/2} h + h^r\right) \), and note that \( \hat{\lambda} \in \Lambda_n \). Then by the usual expansion for the FOC for \( \hat{\lambda} \) we get \( \hat{\lambda} = -\hat{\lambda}^2/\mu_2 + o_p(1) \). Thus expanding \( \hat{P}_\rho \left( f(z), \hat{\lambda} \right) \) we get

\[
\hat{P}_\rho \left( f(z), \hat{\lambda} \right) - \hat{P}_\rho (0) = -2\hat{\lambda}^2\hat{g}_h + \hat{\lambda}^2 \hat{g}_h^2 + o_p(1) = n\hat{g}_h^2/\mu_2 + o_p(1).
\]

Let \( Z = n^{1/2} \left( \hat{g}_h - \mu_h \right)/\mu_2^{1/2} \) and note that \( Z \xrightarrow{d} \mathcal{N}(0, 1) \) so that

\[
\hat{P}_\rho \left( f(z), \hat{\lambda} \right) - \hat{P}_\rho (0) \xrightarrow{a} \left( Z + \mu_h/\mu_2^{1/2} \right)^2 \xrightarrow{d} \chi_1^2
\]

iff \( n^{1/2} \mu_h/\mu_2^{1/2} \to 0 \), which is implied by (III). \( \blacksquare \)
5.2. **LOCAL LINEAR REGRESSION**

5.2 Local linear regression

Let \( z_i' = [y_i, x_i] \) \((i = 1, ..., n)\) be i.i.d. \( \mathbb{R}^2 \)-valued random vectors, and consider the following nonparametric regression

\[
y_i = m(x_i) + \varepsilon_i,
\]

where \( m(\cdot) \) is an unknown regression function, and \( E(\varepsilon_i|x_i) = 0 \), and \( V(\varepsilon_i|x_i) = \sigma^2 < \infty \). One possible way to estimate \( m(\cdot) \) is to use the Nadaraya-Watson or the Gasser-Müller estimators. Here we consider the local linear smoothing estimators - see for example the monograph of Fan & Gijbels (1996).

Let

\[
\hat{m}(x) = \frac{\sum_{i=1}^n W_i y_i}{\sum_{i=1}^n W_i}
\]

denote the local linear estimator for \( m(x) \) where

\[
W_i = K_h(x_i - x) [s_{h2} - (x_i - x) s_{h1}/h],
\]

\[
s_{hl} = \sum_{i=1}^n [(x_i - x)/h]^l K_h(x_i - x)/(nh).
\]

Standard results of Fan & Gijbels (1996) show that

\[
(nh/\beta(x/h))^{1/2} (\hat{m}(x) - m(x))/ (V(x) f(x))^{1/2} \xrightarrow{d} N(B_n, 1)
\]

where \( B_n = \alpha(x/h) (\beta(x/h) V(x)/f(x))^{1/2} m''(x) (nh)^{1/2} h^2 \) with \( V(x) = V(y_i|x_i = x) \), and the functions \( \alpha(\cdot) \) and \( \beta(\cdot) \) involve \( \int_{-1}^{x/h} u^j K(u) du \) \((j = 1, 2, 3)\). By letting \( nh^5 \to 0 \) \( B_n \to 0 \) we obtain a confidence interval with asymptotic correct coverage using the normal approximation with estimated standard error

\[
\beta(x/h) \hat{V}(x)/ (nh \hat{f}(x))^{1/2}
\]

and \( \hat{V}(\cdot), \hat{f}(\cdot) \) are kernel estimates. Alternatively we can use GEL with the same undersmoothing.

Let \( g_{ih} (m(x)) = W_i (y_i - m(x_i)) \) denote the moment indicator, and note that

\[
E [g_{ih} (m(x))] = \gamma(x/h) d^2 m(x) / dx^2 h^2 / 2 + O(h^3)
\]

where \( \gamma(\cdot) \) is a function involving \( \int_{-1}^{x/h} u^j K(u) du \) \((j = 1, 2, 3)\). The following theorem shows that GEL with undersmoothing can be used to obtain confidence intervals for \( m(x) \) at a fixed \( x \) with asymptotic correct coverage. Let

\[
I_\alpha = \left\{ m(x) \mid \hat{P}_\rho \left( m(x), \hat{\lambda} \right) - \hat{P}_\rho (0) \leq c_\alpha \right\}
\]

where \( \Pr(\chi^2_1 \leq c_\alpha) = \alpha \) denote a GEL confidence interval with nominal coverage \( \alpha \).
Theorem 39 Assume that (I) the kernel $K(\cdot)$ is bounded with compact support $[-1,1]$, (II) $f(\cdot)$ (the marginal density of $x_i$), $V(y_i|x_i=\cdot)$ and $m(\cdot)$ have continuous derivatives up to the second order in a neighbourhood of $m$ be used to construct confidence intervals for the Nadaraya-Watson estimator of kernel smoothers to deal with nonparametric problems. Similar arguments could be used effectively in conjunction with GEL.

5.3 Other applications

The above theorems show that GEL can be used effectively in conjunction with kernel smoothers to deal with nonparametric problems. Similar arguments could be used to construct confidence intervals for the Nadaraya-Watson estimator of $m(x)$, or for the derivative of $m(x)$. Another interesting application is with quantile regressions. Whang (2006) shows that EL can be used to obtain valid asymptotic inferences for the slope parameters of quantile regression models.

To elaborate further let

$$g_t(\theta) = x_i [I \{ y_i \leq x'[\theta] - q \}]$$
denote the moment indicator and note that $E [g_i (\theta_0)] = 0$ under the conditional quantile restriction $\Pr (y_i - x'_i \theta_0 \leq 0|x_i) = q$ a.s. Because the indicator function is nondifferentiable, Whang (2006) proposes to use the following smoothed version of $g_i (\cdot)$

$$g_{ih} (\theta) = x_i [K_h (x'_i \theta - y_i) - q].$$

He shows that under regularity conditions similar to those used in the above theorems $2 \sum_{i=1}^n \log \left(1 + \hat{\lambda} g_{ih} (\theta_0)\right) \xrightarrow{d} \chi^2_k$. Using (essentially) the same arguments as those used in Theorem 38 it is possible to generalise his Theorem 2 to GEL. To be specific let $R_\alpha = \left\{ \beta | \left(\hat{P}_\rho (\theta, \hat{\lambda}) - \hat{P}_\rho (0)\right) \leq c_\alpha \right\}$ where $\Pr (\chi^2_k \leq c_\alpha) = \alpha$ denote a GEL confidence region with nominal coverage $\alpha$. Then

$$\Pr (\theta \in R_\alpha) = \alpha + o (1).$$

---

1Theorem 1, which states that the smoothed EL estimator for $\theta$ defined as $\arg \min_{\theta \in \Theta} \sum \log (1 - \lambda' g_{ih} (\theta))$, is also valid for the GEL estimator. It can also be shown that Theorem 3 holds for GEL, under further differentiability assumptions about $\rho (\cdot)$, similar to those described in Section 3.2.
CHAPTER 5. GENERALISED EMPIRICAL LIKELIHOOD AND SMOOTHING
Chapter 6

Generalised empirical likelihood and dependent data

This chapter shows how GEL can be used with dependent observations. We consider two types of dependency structures: martingales and $\alpha$-mixing.

### 6.1 (Generalised) Dual likelihood

Mykland (1995) introduced the concept of dual likelihood (DL henceforth) as device for using likelihood methods in the context of martingale inference. As noted by Mykland (1995) “Martingales methods are a powerful tools for dependent variables inference...A major weakness, however, is that for small samples the quality of the approximation in the martingale CLTs can be quite poor”. DL tries to overcome this problem by introducing a likelihood ratio statistic as an inferential tool for martingales.

The type of martingale we consider in this section is that of a compensated sum of jumps

$$m_t(\theta) = \sum_{0 \leq s \leq t} \Delta m_s(\theta) - \Lambda_t(\theta).$$

(6.1)

Note that (6.1) assumes that $m(\theta)$ does not have infinite total variation. We now give two examples of dependent data structures that can be cast in (6.1)

**Example 40 (Mykland (1995))** (I) Survival data. Let $H(t) = \int_0^t Y(s) d\Lambda_s$,  

47
which specialises to (up to a sign reversal) (log) DL

To estimate (and/or to make inference about) \( m_t \) one can use the martingale

\[
m_t(\theta) = \int_0^t (Y(s)' Y(s))^{-1} Y(s)' dN_t - \Lambda_t
\]

where \( N_t \) is an \( n \)-dimensional vector that jumps from 0 to 1 when a patient dies.

II) Time series. Let \( y_i = \theta [y_{i-1} \ldots y_{i-p}]' + \varepsilon_i \). Then to estimate \( \theta \) one can use the martingale

\[
m_t(\theta) = [y_{i-1} \ldots y_{i-p}]' (y_t - \theta [y_t \ldots y_{i-p}]).
\]

Note that in this case the compensator \( \Lambda_t = 0 \).

We now introduce the notion of generalised DL (GDL henceforth). The latter is based on the following criterion

\[
\hat{P}_\rho^d(\lambda, \theta) = -\lambda' \Lambda_t(\theta) + \sum_{1 \leq s \leq t} \rho \left( \lambda' \Delta m_s(\theta) \right)
\]

which specialises to (up to a sign reversal) (log) DL

\[
(\lambda, \theta) = -\lambda' \Lambda_t(\theta) + \sum_{1 \leq s \leq t} \log \left( 1 - \lambda' \Delta m_s(\theta) \right)
\]

for \( \rho(v) = \log(1 - v) \). Thus GDL is a concave transformation of a martingale.

The following theorem shows that \( \hat{P}_\rho^d(\lambda, \theta) \) can be used in the usual way to obtain asymptotically valid inference for the hypothesis \( H_0 : \theta = \theta_0 \). Let \( \hat{\lambda} = \sup_{\lambda \in \Delta(\theta_0)} \hat{P}_\rho^d(\lambda, \theta) \) where \( \Delta(\theta_0) \) is an open neighbourhood of 0, and let \( m_t^\theta(\theta) \) denote a \( k \)-dimensional triangular array of martingales. Let \( \{m^n(\theta), m^n(\theta)\}_{t_n} = \sum_{0 \leq t \leq t_n} \Delta m_t^\theta(\theta) \Delta m_t^\theta(\theta)' \) denote the quadratic variation of \( m_t^\theta(\theta) \) and let \( \sigma_n = \sigma_{\min}\{m^n(\theta), m^n(\theta)\}_{t_n} \). The following theorem generalises Theorem 1 of Mykland (1995).

**Theorem 41** Assume that (I) \( S_n = m_{t_n}^n(\theta_0)' [m^n(\theta_0), m^n(\theta_0)]_{t_n} m_{t_n}^n(\theta_0) \) is tight, (II) \( \sup_{0 \leq t \leq t_n} \|\Delta m_t^\theta(\theta_0)\| / \sigma_n = o_p(1) \). Then

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

**Proof.** Let \( [m^n(\theta), m^n(\theta)]_{t_n} = Q_n D_n Q_n \) denote the spectral decomposition of the quadratic variation of \( m_t^\theta(\theta) \), and \( U_n(\theta_0) = D_n^{-1/2} Q_n m_{t_n}^n(\theta_0) \). Note that for \( \|\lambda\| < 1/\nu_n = \sigma_n^{1/2} / \left( \sup_{0 \leq t \leq t_n} \|\Delta m_t^\theta(\theta_0)\| \right) \) an expansion about 0 and T give

\[
\left| \hat{P}_\rho^d(\lambda, \theta_0) - \hat{P}_\rho^d(0) + \lambda' U_n(\theta_0) + \lambda' \lambda/2 \right| \leq (\rho_2(\|\lambda\| \nu_n) - 1) \lambda' \lambda/2 \leq o_p(1)
\]
hence $\hat{\lambda}$ is consistent, and by concavity of $\tilde{P}_d^\rho (\cdot)$ is tight because of (I). Then

$$2 \left( \tilde{P}_d^\rho (\lambda, \theta_0) - \rho (0) \right) = U_n (\theta_0)' U_n (\theta_0) + o_p (1)$$

and the result follows by Theorem 5.1 of Helland (1982) and CMT. 

A situation of particular practical relevance and theoretical interest that does not fit into this framework is that of unstable autoregressive processes. Next section shows how to adapt GEL to deal with these type of processes.

### 6.1.1 GDL for unstable autoregressive processes

In this section we show how GDL can be used to obtain valid asymptotic inference in the context of autoregressive models with possibly multiple and/or complex unit roots. To be specific we show that the asymptotic distributions of GDL statistic both under the null and under the general sequence of local alternative hypotheses considered by Jeganathan (1991) are linear combinations of quadratic forms of certain stochastic integrals involving Brownian motion.

Let $\theta_0 = \begin{bmatrix} \theta_{10} & \cdots & \theta_{p0} \end{bmatrix}'$ denote the vector of unknown parameters of the autoregressive process of order $p$ (AR(p))

$$z_t = \theta_0' z_{t-1} + \varepsilon_t \quad t = 1, ..., n$$

(6.2)

where $z_t = \begin{bmatrix} z_t & \cdots & z_{t-p+1} \end{bmatrix}'$, and the innovations $\varepsilon_t$ form a martingale difference sequence with respect to the natural filtration $\mathcal{F}_t$, that is $\varepsilon_t$ is $\mathcal{F}_t$-measurable with $E(\varepsilon_t | \mathcal{F}_{t-1}) = 0$ a.s. for every $t$. As mentioned in Example 7 inference for $\theta$ in (6.2) can be based on

$$m_t (\theta) = (z_t - \theta' z_{t-1}) z_{t-1},$$

which is a martingale for $\theta_0$. It is well known that the distribution of the least squares estimator of $\theta_0$ (and related test statistics) depends on the location of the roots of the characteristic polynomial of (6.2), that is

$$\Phi_p (z) = 1 - \sum_{j=1}^{p} \theta_j z^j$$

(6.3)

In what follows we often use $B$ in place of $B (r)$ to denote a standard Brownian motion on $C [0, 1]$, the space of continuous functions on the unit interval $[0, 1]$, and suppress the Lebesgue measure $dr$ in integrals so that $\int_{0}^{1} B (r) \, dr$ becomes simply $\int_{0}^{1} B$. The following theorem establishes the weak convergence in the Skorohod space $D [0, 1]^p$ of the DGEL statistic for the hypothesis $H : \theta = \theta_0$ for unstable autoregressive processes.
CHAPTER 6. GENERALISED EMPIRICAL LIKELIHOOD AND DEPENDENT DATA

Theorem 42 Assume that (I) the initial values \(y_0, ..., y_{1-p}\) are fixed \(\mathcal{F}_0\)-measurable and are \(O_p(1)\), (II) \(E(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma^2\) a.s for every \(t\), \(\sup_t E(\varepsilon_t^{2+\alpha} |\mathcal{F}_{t-1}) < \infty\) for some \(\alpha > 0\). Then if all of the roots of (6.3) are either outside or on the unit circle, with at least one unit root on the unit circle

\[
2 \left( \hat{P}^d (\lambda, \theta_0) - \rho(0) \right) \xrightarrow{w} \Xi' \Delta^{-1} \Xi,
\]

where

\[
\Xi = \begin{bmatrix} \int_0^1 \zeta dB & \int_0^1 \eta'dB & \int_0^1 \zeta_1 dB & \cdots & \int_0^1 \zeta dB \end{bmatrix}' , \quad \Delta = \text{diag} \begin{bmatrix} \int_0^1 F & \int_0^1 \bar{F} & \int_0^1 H_1 & \cdots & \int_0^1 H_1 \end{bmatrix} ,
\]

and the random elements in \(\Xi, \Delta\) are as defined in Theorem 3.5.1 of Chan & Wei (1988).

Proof. As in Chan & Wei (1988) we use a suitable normalising matrix, say \(N_n\), to decompose the original AR\((p)\) process \(y_t\) into several separate processes according to the nature of the roots of (4.4), so that for \(p' = a + b + 2 \sum_{k=1}^p d_k \leq p\)

\[
N_n y_t = Y_{nt} = \begin{bmatrix} u_t(1)/n & \cdots & u_t(a)/n^a & v_t(1)/n & \cdots & v_t(b)/n^b & x_t(d_1)/n & \cdots & x_t(d_l)/n^{d_l} & z_t/n^{1/2} & \cdots & z_{t-q+1}/n^{1/2} \end{bmatrix}'
\]

(6.5)

where \(u_t(j)/n^j = (1 - L)^{-j} \varepsilon_t (j = 1, ..., a), v_t(j)/n^j = (1 + L)^{-j} \varepsilon_t (j = 1, ..., b), (1 - 2 \cos \theta_k L + L^2)^{d_k} x_t(d_k)(j) = \varepsilon_t\) for \((j = 1, ..., d_k)\)

\[
x_t(d_j) = \begin{bmatrix} x_t(d_k)(1)/n & x_{t-1}(d_k)(1)/n & \cdots & x_t(d_k)(d_k)/n^{d_k} \\
x_{t-1}(d_k)(d_k)/n^{d_k} \end{bmatrix}'.
\]

and \(\Psi(B) z_t = \varepsilon_t\) with \(\Psi(B)\) is a polynomial of degree \(q = p - p'\) with all the roots outside the unit circle. Let \(\Lambda_n = \{ \lambda : \| N_n^{-1} \lambda \| \leq C \}\) for some \(C > 0\). By results of Chan & Wei (1988) \(\sum Y_{nt} \varepsilon_t\) and \(\sum Y_{nt} Y_{nt}^t\) are bounded in probability; furthermore by Chuang & Chan (2002) \(\| \sum Y_{nt} Y_{nt}^t (\varepsilon_t^2 - \sigma^2) \| = o_p(1)\), and max \(\| Y_{nt} \varepsilon_t \| = o_{a.s.}(1)\) and thus

\[
\max_t \sup_{\lambda \in \Lambda_n} \| \lambda' (N_n^{-1})^t Y_{nt} \varepsilon_t \| = o_p(1).
\]
6.1. (GENERALISED) DUAL LIKELIHOOD

Then by the usual Taylor expansion about \(0\)
\[
\sup_{\lambda \in \Lambda_n} \tilde{P}_\rho (\theta_0, \lambda) \leq \rho_0 - \lambda' \left( N_n^{-1} \right)' \sum_t Y_{nt} \varepsilon_t + \\
\lambda' \left( N_n^{-1} \right)' \sum_{i=1}^n \rho_2 \left( \lambda' \left( N_n^{-1} \right)' Y_{nt} \varepsilon_t \right) Y_{nt} Y_{nt}^t N_n^{-1} \lambda / 2
\]
\[
\leq \rho_0 - \left\| N_n^{-1} \lambda \right\| \left\| \sum_t Y_{nt} \varepsilon_t \right\| - \zeta_n \left\| N_n^{-1} \lambda \right\|^2 / 4,
\]
where \(\zeta_n > 0\) a.s. Then \(\left\| N_n^{-1} \lambda \right\| \leq \left\| \sum_t Y_{nt} \varepsilon_t \right\| = O_p(1)\). To find a stochastic approximation of \(N_n^{-1} \lambda\) we expand the FOCs about \(0\) and use CMT to obtain
\[
0 = \sum \rho_1 \left( \left( N_n^{-1} \lambda \right)' Y_{nt} \varepsilon_t \right) N_n^{-1} Y_{nt} \varepsilon_t = -N_n^{-1} \sum Y_{nt} \varepsilon_t - \sum_{i=1}^n N_n^{-1} Y_{nt} Y_{nt}^t N_n^{-1} \lambda + o_p(1)
\]
which yields
\[
N_n^{-1} \lambda = \left( \sum Y_{nt} Y_{nt}^t \right)^{-1} \sum Y_{nt} \varepsilon_t + o_p(1) \tag{6.6}
\]
since by results of Chan & Wei (1988) \(\sum Y_{nt} Y_{nt}^t\) is nonsingular a.s.. The conclusion of the theorem follows by the usual second order expansion about \(0\) and (6.6) which shows that
\[
2 \left( \tilde{P}_\rho^d (\lambda, \theta_0) - \rho (0) \right) = \sum Y_{nt} \varepsilon_t' \left( \sum Y_{nt} Y_{nt}^t \right)^{-1} \sum Y_{nt} \varepsilon_t + o_p(1)
\]
and by the results of Chan & Wei (1988).

To evaluate the power properties of the DGEL statistic we consider the same type of (local) alternative hypotheses considered by Jeganathan (1991), that is \(H_n : \theta_n = \theta_0 + N_n \delta_n\), where \(N_n\) is a suitable normalising matrix (see (6.11) in the Appendix), and \(\delta_n\) is a sequence of (nonrandom) vectors such that \(\delta_n \rightarrow \delta_0\) for \(n \rightarrow \infty\) and \(0 < \delta_0' \delta_0 < \infty\). For example, in the case of an unstable AR(1) process \(H_n : \theta_n = 1 + n^{-1} \delta_n\) corresponds to the sequence of local-to-unity alternatives considered in detail by Chan & Wei (1987) and Phillips (1987).

**Proposition 43** Under the same assumptions of Theorem 42, and under \(H_n : \beta_n = \beta_0 + N_n \delta_n\), if \(\delta_n \rightarrow \delta_0 \equiv \left[ u_0 \ v_0 \ \cdots \ u_l \ v_l \ z \right]\), then
\[
2 \left( \tilde{P}_\rho^d (\lambda, \theta_n) - \rho (0) \right) \xrightarrow{w} \Xi_{\delta_0} \Delta_{\delta_0}^{-1} \Xi_{\delta_0},
\]
where
\[
\Xi_{\delta_0} = \left[ \int_0^1 \xi_1' (u_0) dB \ \int_0^1 \xi_1' (v_0) dB \ \int_0^1 \xi_1' (u_1, v_1) dB \ \cdots \ \int_0^1 \xi_1' (u_l, v_l) dB \ \cdots \ \int_0^1 H (u_l, v_l) \ \Gamma_3 \right]',
\]
\[
\Delta_{\delta_0} = \text{diag} \left[ \int_0^1 \Gamma_1 (u_0) \ \int_0^1 \Gamma_2 (v_0) \ \int_0^1 H (u_1, v_1) \ \cdots \ \int_0^1 H (u_l, v_l) \ \Gamma_3 \right],
\]
and the random elements in $\Xi_{\delta_0}$, $\Delta_{\delta_0}$ are as defined in Theorem 1 of Jeganathan (1991).

**Proof.** See Bravo (2006) (for the ECR statistic)

GDL can be used to obtain valid inferences in unstable autoregressive models that contain deterministic components. Consider

$$z_t' = \eta_0 d_t + \varepsilon_t$$

$$z_t = \theta_0' z_{t-1} + \varepsilon_t \quad t = 1, \ldots, n,$$

where $z_t = [z_t \cdots z_{t-p+1}]'$ and $d_t = [1 \cdots t^{q-1}]'$ is a polynomial trend$^1$.

Let $h = [h_1(r) \cdots h_q(r)]'$, $H = \int_0^1 hh'$, and for a random vector $X \in D [0, 1]^q$, let $X_h$ denote its projection on the space orthogonal to $(\int_0^1 h')H^{-1}h$. Let $2 \left( \tilde{P}_d^d (\lambda, \theta_0, \eta) - \rho (0) \right)$ denote the profile GDL statistic.

**Proposition 44** Under the same assumptions of Theorem 42, and under $H_0 : \theta = \theta_0$, then

$$2 \left( \tilde{P}_d^d (\lambda, \theta_0, \eta) - \rho (0) \right) \overset{w}{\to} \Xi_h' \left( \Delta - \Upsilon H^{-1} \Upsilon' \right)^{-1} \Xi_h,$$

where

$$\Xi_h = \left[ \int_0^1 \xi_d dB \quad \int_0^1 \eta_d dB \quad \int_0^1 \zeta_1 dB \quad \cdots \quad \int_0^1 \zeta_l dB \quad N_h (0, \Sigma) \right]'$$

$$\Upsilon = \sigma \int_0^1 \left[ \xi dB \quad \eta dB \quad H_1 dB \quad \cdots \quad H_l dB \quad \Sigma \right] h',$$

and $\Delta$ is as in (6.4).

**Proof.** See Bravo (2006) (for the ECR statistic)

**Remark 45** The computation of the profile statistic $\tilde{P}_d^d (\lambda, \theta_0, \eta)$ can be carried out using the residuals of the preliminary regression $z_t'$ on $d_t$ and simply proceed with the analysis as in the case without the deterministic components.

**Example 46** (Double unit roots with intercept) We consider an $AR (2)$ process with an intercept

$$z_t = \sum_{j=1}^2 \theta_j z_{t-j} + \eta_0 + \varepsilon_t, \quad t = 1, \ldots, n.$$  \hfill (6.7)

$^1$In fact the result holds for deterministic sequences of general form satisfying the following regularity condition: for $j = 1, \ldots, q$ there exist $\lambda_j$ and linearly independent bounded functions $h_j (r)$ such that $d_j (n) / n^{\lambda_j} \to h_j (r)$ as $n \to \infty$ uniformly in $r \in [0, 1]$.  

The null hypothesis of interest is that of double (positive) unit roots in the characteristic polynomial \( \Phi_2(z) = 1 - \sum_{j=1}^{2} \theta_j z^j \) of (6.7), that is under \( H_0 \), \( z_{1,2} = 1 \). This hypothesis can be tested using the profile GDL statistic \( \tilde{P}_\rho^d(\lambda, \theta_0, \tilde{\eta}) \) with \( \theta_0 = \begin{bmatrix} 2 & -1 \end{bmatrix} \). By Proposition 44

\[
2 \left( \tilde{P}_\rho^d(\lambda, \theta_0, \tilde{\eta}) - \rho(0) \right) \xrightarrow{w} \left[ \int_0^1 \overline{B}_m dB \quad \int_0^1 B_m dB \right] F_m^{-1} \left[ \int_0^1 \overline{B}_m dB \quad \int_0^1 B_m dB \right]',
\]

where

\[
F_m = \begin{bmatrix}
\int_0^1 B_m^2 \\
\int_0^1 \overline{B}_m B_m \\
\int_0^1 \overline{B}_m \\
\int_0^1 B_m^2
\end{bmatrix},
\]

and \( B_m \) and \( \overline{B}_m \) are, respectively, a demeaned and an integrated demeaned Brownian motions. The following table report the finite sample size and power of the Euclidean (\( \rho = EU \)), exponential (\( \rho = ET \)), empirical (\( \rho = EL \)) likelihood and of the F-type statistic \( \Phi_2(2) \) suggested by Hasza & Fuller (1979) for sample sizes \( n = 50 \) and \( n = 100 \). As with the previous example, size and power calculations were based on 5000 and 1000 replications, respectively, using the critical values obtained from the Monte Carlo distribution of (6.8) using the demeaned time series \( z_t - \sum_{i=1}^{n} z_t / n \). For the \( \Phi_2(2) \) statistic we used the critical
values tabulated by Hasza & Fuller (1979).

Finite sample size and power of the GDL test statistic for double unit roots at 0.10 and 0.05 nominal level

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\rho = EU$</th>
<th>$\rho = ET$</th>
<th>$\rho = EL$</th>
<th>$\Phi_2 (2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.01</td>
<td>1.01</td>
<td>0.115</td>
<td>0.126</td>
<td>0.135</td>
<td>0.115</td>
</tr>
<tr>
<td>2.01</td>
<td>1.00</td>
<td>0.903</td>
<td>0.907</td>
<td>0.917</td>
<td>0.898</td>
</tr>
<tr>
<td>2.01</td>
<td>0.99</td>
<td>0.982</td>
<td>0.984</td>
<td>0.986</td>
<td>0.981</td>
</tr>
<tr>
<td>2.00</td>
<td>1.01</td>
<td>0.434</td>
<td>0.465</td>
<td>0.487</td>
<td>0.354</td>
</tr>
<tr>
<td>2.00</td>
<td>1.00</td>
<td>0.114</td>
<td>0.103</td>
<td>0.102</td>
<td>0.112</td>
</tr>
<tr>
<td>2.00</td>
<td>0.99</td>
<td>0.879</td>
<td>0.855</td>
<td>0.865</td>
<td>0.869</td>
</tr>
<tr>
<td>1.99</td>
<td>1.01</td>
<td>0.700</td>
<td>0.712</td>
<td>0.721</td>
<td>0.605</td>
</tr>
<tr>
<td>1.99</td>
<td>1.00</td>
<td>0.401</td>
<td>0.420</td>
<td>0.438</td>
<td>0.292</td>
</tr>
<tr>
<td>1.99</td>
<td>0.99</td>
<td>0.257</td>
<td>0.280</td>
<td>0.297</td>
<td>0.165</td>
</tr>
<tr>
<td>$n = 100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.01</td>
<td>1.01</td>
<td>0.168</td>
<td>0.174</td>
<td>0.170</td>
<td>0.124</td>
</tr>
<tr>
<td>2.01</td>
<td>1.00</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2.01</td>
<td>0.99</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>2.00</td>
<td>1.01</td>
<td>0.886</td>
<td>0.945</td>
<td>0.956</td>
<td>0.881</td>
</tr>
<tr>
<td>2.00</td>
<td>1.00</td>
<td>0.100</td>
<td>0.107</td>
<td>0.102</td>
<td>0.103</td>
</tr>
<tr>
<td>2.00</td>
<td>0.99</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.996</td>
</tr>
<tr>
<td>1.99</td>
<td>1.01</td>
<td>0.989</td>
<td>0.993</td>
<td>0.992</td>
<td>0.994</td>
</tr>
<tr>
<td>1.99</td>
<td>1.00</td>
<td>0.862</td>
<td>0.932</td>
<td>0.932</td>
<td>0.859</td>
</tr>
<tr>
<td>1.99</td>
<td>0.99</td>
<td>0.299</td>
<td>0.305</td>
<td>0.321</td>
<td>0.245</td>
</tr>
</tbody>
</table>

6.2 Blocking and smoothing methods

The proof of Theorem 15 shows that the GEL estimator $\hat{\theta}$ is still $n^{1/2}$-consistent and asymptotically normal even with weakly dependent observations, but is less efficient than the efficient GMM estimator. More importantly GEL test statistics are no longer asymptotically chi-squared distributed\(^2\). One way to solve this problem is to consider blocking techniques as suggested by Kitamura (1997b). Alternatively one can use kernel smoothing techniques as suggested by Kitamura & Stutzer (1997), Smith (1997) and Smith (2004), among others. We

\(^2\)For the exactly identified case it is possible to show that the GEL test statistic for $H_0 : \theta = \theta_0$ has asymptotic distribution $\sum_{j=1}^{L} \omega_j \lambda_{j}^2$ where the $\omega_j$ are the eigenvalues of $[E[g(\theta_0)g(\theta_0)^T]]^{-1} \sum_{j=1}^{L} COV [g_j(\theta_0)g_j(\theta_0)]$. 
first briefly consider the latter. Let

\[ g_{st}(\theta) = \sum_{s=t-n}^{t-1} K_h(s) g_{t-s}(\theta) / (nh) \]

denote a smoothed moment indicator where \( K_h(\cdot) \) is a kernel function. The kernel weights \( K_h(s) / h \) are in the same spirit of those used in the heteroskedasticity autocorrelation consistent (HAC) covariance matrix estimation; see for example Newey & West (1987) and Andrews (1991).

To be specific we can define a smoothed GEL criterion function as

\[ \hat{P}_b^\lambda (\theta, \lambda) := \sum_{i=1}^{n} \rho \left( k' g_{st}(\theta) \right) / n \]  

(6.9)

where \( k \) is a normalisation constant that has no effect on the GEL estimator but makes the scale of the auxiliary estimator \( \hat{\lambda} \) comparable for different choices of kernels. Smith (2004) shows that (6.9) can be used to obtain efficient estimators and asymptotically valid inference.

An alternative (asymptotically equivalent) approach is to consider blocking techniques, which are also used in the bootstrap literature, see for example Politis & Romano (1992). The basic idea is to construct “new” observations by considering blocks of the original observations, and base estimation and inference on the resulting sequence of blocks. This procedure preserves nonparametrically the dependent structure of the data, delivering therefore valid asymptotic inference. As in Kitamura (1997b), let \( l = l(n) \) and \( m = m(n) \) denote two integers functions of \( n \) such that \( 1 \leq l \leq m \), and \( \lim_{n \to \infty} m = \infty \). For each \( i \in \mathbb{N} \), let \( b_{i,m,l} = \left[ z_{i-1}^{(l+1)} \ldots z_{i-1}^{(l+m)} \right]' \) be a block of \( m \) consecutive observations starting from \((i - 1) l + 1\). Note that \( m \) is the block length and \( l \) is the separation between block starting points. Thus if \( l = m \) the resulting sequence of blocks is nonoverlapping, while if \( l = 1 \) it is fully overlapping. Define now the blockwise moment function

\[ \psi(b_{i,m,l}, \theta) := \psi_i(\theta) = \sum_{j=1}^{m} g \left( z_{i-1}^{(l+j)} , \theta \right) / m, \]

and note that if (2.1) holds then \( E \left[ \psi_i(\theta_0) \right] = 0 \ \forall i \). Blockwise GEL (BGEL henceforth) estimation and inference for \( \theta_0 \) is based on the BGEL criterion function

\[ \widehat{P}_b^\lambda (\theta, \lambda) := \sum_{i=1}^{q} \rho \left( X_i \psi_i(\theta) \right) / q \]
where \( q = \lfloor (n - m) / l + 1 \rfloor \) is the total number of blocks and \(|\cdot|\) is the integer part function. The following theorem generalises Theorem 15 to weakly dependent (possibly nonstationary) observations.

**Theorem 47** Assume that (I) \( z_t \) is a strong mixing sequence of size \(-2\alpha / (\alpha - 2)\) where \( \alpha > 2 \), (II) The parameter space \( \Theta \) is compact, (III) (a) \( E \left( \sup_{t} \sup_{\theta \in \Theta} \| g_t (\theta) \|^{2\alpha + \delta} \right) < \infty \) for some \( \delta > 0 \), (b) \( \lim_{n \to \infty} E \left[ \tilde{g} (\theta) \right] \) exists uniformly in \( \Theta \) and is continuous, (IV) \( \lim_{n \to \infty} \sum_{t=1}^{n} E \sup_{\theta \in \Theta} \| g_t (\theta) \| / n < \infty \), (V) \( \Omega (\theta) := \lim_{n \to \infty} V \left[ n^{1/2} \tilde{g} (\theta) \right] \) exists and it is positive definite uniformly in \( \Theta \), (VI) \( \rho (\cdot) \) is twice continuously differentiable in an open neighbourhood of 0, (VII) (a) \( G_t (\theta) \) is twice-continuously differentiable in a convex neighbourhood \( G \) of \( \theta_0 \) \forall t \) \( E \sup_{\theta \in \Theta} \| g_t (\theta) \| \) and \( \lim_{n \to \infty} E \left[ \sum_{t=1}^{n} \sum_{j=1}^{l} \right. \left. g_t (\theta) / \partial \theta_j \right] = \frac{1}{n} \) \( \left. \sum_{t=1}^{n} \sum_{j=1}^{l} \right. \left. g_t (\theta) / \partial \theta_j \right] \) exist uniformly in \( \Theta \) and are continuous, (viii) \( \lim_{n \to \infty} \sum_{t=1}^{n} E \sup_{\beta \in \Theta} \right\| \partial^k g_t (\theta) / \partial \theta_j \right\| / n < \infty \), (v) \( \right. \right. \lim_{n \to \infty} E \left[ \tilde{g} (\theta_0) \right] = 0 \) where \( G_0 = \lim_{n \to \infty} E \left[ \tilde{G} (\theta_0) \right] \), (vi) \( G (\beta_0) \Omega (\beta_0)^{-1} G (\beta_0) \) is nonsingular. \( \right. \right. \)

Then

**Theorem 48** Assume A1-A5 hold. Then for \( m = o \left( n^{1/2} \right) \)

\[
\begin{bmatrix}
(n^{1/2}/m) \tilde{\lambda} \\
n^{1/2} (\tilde{\theta} - \theta_0)
\end{bmatrix} \xrightarrow{d} N \left( 0, \begin{bmatrix} \Psi_0 & 0 \\ 0 & (G_0^* \Omega_0^{-1} G_0)^{-1} \end{bmatrix} \right),
\]

where \( \Psi (\beta_0) \) is as in (2.3) of Theorem 15.

**Proof.** The proof is similar to that of Theorem 15 and uses ULLN as in Andrews (1987) applied to \( \psi (\theta) \) and its first derivative, and LLN and CLT for strongly mixing processes to show that \( \| \tilde{\psi} (\tilde{\theta}) \| = O_p (n^{-1/2}) \) and \( \| \tilde{\lambda} \| = O_p (m/n^{1/2}) \). Then by mean value expansion

\[
m \left\| \sum_{t=1}^{t} \left( \partial^k \psi (\theta) / \partial \theta_j \right) \psi (\theta) / q + \sum_{t=1}^{t} \left( \partial^k \psi (\theta) / \partial \theta_j \right) \psi (\theta) / q \right\| \leq \left( \sup_{\theta \in \Theta} \| \partial^k g (\theta) / \partial \theta_j \| \right)^{1/\alpha} / n \left( \sum_{t=1}^{t} \left( \sup_{\theta \in \Theta} \| \partial^k+1 g (\theta) / \partial \theta_j \| / n \right)^{\alpha-1/\alpha} \right) \times \left( m / n^{1/2} \right) n^{1/2} \| \tilde{\theta} - \theta_0 \| + O_p (m^2 / ln) = o_p (1) \right. \text{ for } k = 0, 1,
\]

\[
m \left\| \sum_{j=1}^{t} \psi_{ij} (\tilde{\theta}) \partial \psi (\tilde{\theta}) / \partial \theta_j \right\| = o_p (1)
\]

and therefore by CMT and \( \| \tilde{\lambda} \| = o_p (1) \| \sum_{t=1}^{t} \rho (\tilde{\lambda} \psi (\tilde{\theta}) \partial \psi (\tilde{\theta}) / \partial \theta_j) \right\| = o_p (1) \]
6.2. BLOCKING AND SMOOTHING METHODS

Furthermore

\[
\sum \rho_1 \left( \hat{X} \psi_i (\overline{\theta}) \right) \partial \psi_i (\overline{\theta}) / \partial \theta / q + G (\theta_0) = o_p (1),
\]

\[
\sum \sum_{j=1}^t \rho_2 \left( \hat{X} \psi_i (\overline{\theta}) \right) X_j \partial^2 \psi_i (\overline{\theta}) / \partial \theta \partial \theta_j / q = o_p (1),
\]

and as in the proof of (2.15) of Theorem 15

\[
n^{-1} \left[ m \partial^2 W (\hat{X}, \overline{\theta}) / \partial \lambda \partial \lambda' \quad \partial^2 W (\hat{X}, \overline{\theta}) / \partial \theta \partial \lambda' \right] \rightarrow - \left[ \Omega \quad G \right] = M_0.
\]

Then

\[
n^{1/2} \left[ \hat{X}' / m, (\hat{\theta} - \theta_0) \right]' = \left[ \Psi (\beta_0), N (\beta_0)' \right] n^{1/2} \hat{\psi} (\beta_0) + o_p (1),
\]

and the conclusion follows by standard arguments (see Bravo (2007) for more details).

BGEL based inference for \( \theta \) can be obtained using the same type of test statistics used for example in Theorem 10. Here we briefly consider specifications tests based on additional moment conditions as originally developed by Newey (1985). Let \( \beta = [\alpha', \beta]' \) where \( \alpha \) is an \( \mathbb{R}^p \)-valued vector of additional parameters, and suppose that there exists an \( \mathbb{R}^s \)-valued \( (s \leq p) \) vector of functions \( h (z_t, \beta) := h_t (\beta) \) satisfying

\[
E [h_t (\beta_0)] = 0, \quad \forall t \quad (6.10)
\]

The information contained in this additional set of moment conditions can naturally be incorporated into BGEL estimation. To be specific let \( l_t (\beta) = [g_t (\hat{\theta})', h_t (\beta)'] \) denote the “augmented” moment indicator, and let \( \psi_i^a (\beta) = \sum_{j=1}^m l_{(i-1)t+j} (\beta) / m \) denote the blockwise version of \( l (\cdot) \). Finally let

\[
\hat{P}_p^b (\beta, \lambda, \varphi) = \sum_{i=1}^q \rho (\mu' \psi_i^a (\beta)) / q
\]

where \( \mu = [\lambda', \varphi]' \) and \( \varphi \) is an \( \mathbb{R}^s \)-valued vector of unknown auxiliary parameters associated with \( h_t (\beta) \). Test statistics for the additional moment conditions (6.10) may be constructed by imposing the restriction \( \varphi = 0 \) into the estimation of \( \hat{P}_p^b (\beta, \mu) \).

\[
D^\rho = 2c_n (\hat{P}_p^b (\beta, \mu) - \hat{P}_p^b (\check{\beta}^c, \check{\mu}^c)),
\]

\[
LM^\rho = (n/m^2) (\hat{\varphi}')' \left[ S_{\hat{\varphi}} \Delta (\beta) S_{\hat{\varphi}}' \right]^{-1} \hat{\varphi}^c,
\]

\[
S^\rho = \sum_{i=1}^q s_i (\beta)' / q^{1/2} S_{\varphi} \Delta (\beta) S_{\varphi}^{1/2} \sum_{i=1}^q s_i (\beta) / q^{1/2}
\]
where \( c_n = (q/mn) \) is a correction factor that accounts for the overlap in the blocks,

\[
\Delta(\beta) = \Xi(\beta)^{-1} \left( I - L(\beta) \Gamma(\beta)^{-1} L(\beta)' \Xi(\beta)^{-1} \right)
\]

\[
s_i(\beta) = \rho_1 \left( \tilde{\lambda} \psi_i(\beta) \right) \psi_i(\beta),
\]

and \( S^\varphi = [0, I] \) is a selection matrix such that \( S^\varphi \mu = \varphi \) is then possible.

**Theorem 49** Assume that the assumptions of Theorem 47 hold for \( h(\beta) \) with parameter space \( B = A \times \Theta \) is compact. Then under (6.10)

\[
D^\rho, LM^\rho, MC^\rho, S^\rho \overset{d}{\rightarrow} \chi^2(s).
\]

**Proof.** See Bravo (2007) for details.

**Remark 50** The score statistic \( S^\rho \) can be interpreted as the blockwise version of Newey’s (1985) specification test. Some algebra shows that

\[
S^\rho = n \left[ \hat{t}(\theta_0)' \Delta(\theta_0) \hat{t}(\theta_0) - \hat{\theta}(\theta_0)' \Psi(\theta_0) \hat{\theta}(\theta_0) \right] + o_p(1),
\]

which is in fact the asymptotic approximation of

\[
n \left[ \hat{t}(\bar{\theta})' \Xi(\bar{\theta})^{-1} \hat{t}(\bar{\theta}) - \hat{\theta}(\bar{\theta})' \Omega(\bar{\theta})^{-1} \hat{\theta}(\bar{\theta}) \right]
\]

i.e. the difference of two quadratic forms in the estimated sample moments as suggested by Newey (1985).

**Example 51** We consider as in Andersen & Sorensen (1996) GMM estimation of the stochastic volatility model

\[
y_t = \sigma_t u_t
\]

\[
\ln \sigma_t^2 = \beta_1 + \beta_2 \ln \sigma_{t-1}^2 + \beta_3 \varepsilon_t
\]

where \([u, \varepsilon_t]' \sim N(0, I)\. The moment condition is

\[
E \begin{bmatrix}
|y_t| - (2/\pi)^{1/2} \exp(\mu/2 + \sigma^2/8) \\
y_t^2 - \exp(\mu + \sigma^2/2) \\
|y_t y_{t-j}| - (2/\pi)^{1/2} \exp(\mu + \sigma^2/2) \exp(\beta_2' \sigma^2/4)
\end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
\]

for \( j = 1, 3, 5 \). where \( \mu = \beta_1 / (1 - \beta_2) \) and \( \sigma^2 = \beta_2^2 / (1 - \beta_2^2) \). In the simulations we use \( \beta = [-.368, .95, .26]' \). The degree of overidentification is 2. In
each example we compute two versions of the statistics $D^\rho$, $LM^\rho$ and $S^\rho$ and in (2.7): one based on the the exponential tilting ($\rho = ET$) and the other based on the Euclidean distance ($\rho = CU$). In addition we compute the $LM^\rho$ statistic using the (blockwise) implied probabilities defined as

$$\hat{\pi}_i = \rho_1 \left( \hat{\chi}' \psi_i \left( \hat{\theta} \right) \right) / \sum_{j=1}^q \rho_1 \left( \hat{\chi}' \psi_j \left( \hat{\theta} \right) \right),$$

(6.11)

and we also consider the prewhitened analogue of the proposed test statistics. These statistics are compared to

$$J_n \left( \hat{\theta}_{GMM} \right) = n \tilde{g} \left( \hat{\theta}_{GMM} \right)' \hat{\Omega} \left( \hat{\theta}_{GMM} \right)^{-1} \tilde{g} \left( \hat{\theta}_{GMM} \right),$$

with the Newey-West (Newey & West 1987) estimator for $\hat{\Omega} \left( \hat{\theta}_{GMM} \right)$, and the block variance estimator for $\hat{\Omega} \left( \hat{\theta}_{GEL} \right)$ with overlapping blocks (i.e. with $l = 1$). These estimators are asymptotically equivalent for $m = o \left( n^{1/2} \right)$, and have the optimal (lag)-length parameter $m^* = \left[ \gamma n^{1/3} \right]$; for any choice of finite $\gamma > 0$. In the simulations we consider two types of rules for choosing $m$: a fixed one, and a data-dependent one based on Newey & West’s (1994) nonparametric method. For the prewhitening we use simple univariate autoregressions of order one as linear filter, instead of the vector autoregression recommended by Andrews & Monahan (1992). The following table reports the empirical sizes for sample size
CHAPTER 6. GENERALISED EMPIRICAL LIKELIHOOD AND DEPENDENT DATA

\[ n = 1000 \text{ using 5000 replications.} \]

Empirical size for lognormal stochastic volatility model with \( n = 1000 \)

<table>
<thead>
<tr>
<th>( m )</th>
<th>( \alpha )</th>
<th>( J )</th>
<th>( D_{ET}^{\alpha} )</th>
<th>( S_{ET}^{\alpha} )</th>
<th>( LM_{ET}^{\alpha} )</th>
<th>( LM_{ET}^{\alpha}(b) )</th>
<th>( D_{CU}^{\alpha} )</th>
<th>( LM_{CU}^{\alpha} )</th>
<th>( LM_{CU}^{\alpha}(b) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>.100</td>
<td>.214</td>
<td>.146</td>
<td>.168</td>
<td>.241</td>
<td>.160</td>
<td>.222</td>
<td>.240</td>
<td>.138</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.139</td>
<td>.099</td>
<td>.064</td>
<td>.198</td>
<td>.108</td>
<td>.146</td>
<td>.157</td>
<td>.086</td>
</tr>
<tr>
<td></td>
<td>.010</td>
<td>.048</td>
<td>.046</td>
<td>.031</td>
<td>.136</td>
<td>.051</td>
<td>.114</td>
<td>.122</td>
<td>.059</td>
</tr>
<tr>
<td>12</td>
<td>.100</td>
<td>.187</td>
<td>.124</td>
<td>.152</td>
<td>.203</td>
<td>.149</td>
<td>.209</td>
<td>.222</td>
<td>.129</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.119</td>
<td>.084</td>
<td>.058</td>
<td>.176</td>
<td>.100</td>
<td>.137</td>
<td>.145</td>
<td>.081</td>
</tr>
<tr>
<td></td>
<td>.010</td>
<td>.041</td>
<td>.039</td>
<td>.021</td>
<td>.115</td>
<td>.048</td>
<td>.107</td>
<td>.113</td>
<td>.055</td>
</tr>
<tr>
<td>( m^*(c) )</td>
<td>.100</td>
<td>.181</td>
<td>.119</td>
<td>.148</td>
<td>.223</td>
<td>.144</td>
<td>.192</td>
<td>.215</td>
<td>.125</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.115</td>
<td>.080</td>
<td>.056</td>
<td>.193</td>
<td>.097</td>
<td>.126</td>
<td>.140</td>
<td>.078</td>
</tr>
<tr>
<td></td>
<td>.010</td>
<td>.039</td>
<td>.037</td>
<td>.019</td>
<td>.126</td>
<td>.046</td>
<td>.098</td>
<td>.109</td>
<td>.053</td>
</tr>
<tr>
<td>8</td>
<td>.100</td>
<td>.173</td>
<td>.124</td>
<td>.147</td>
<td>.187</td>
<td>.137</td>
<td>.173</td>
<td>.189</td>
<td>.118</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.112</td>
<td>.084</td>
<td>.056</td>
<td>.154</td>
<td>.092</td>
<td>.113</td>
<td>.124</td>
<td>.073</td>
</tr>
<tr>
<td></td>
<td>.010</td>
<td>.038</td>
<td>.039</td>
<td>.009</td>
<td>.106</td>
<td>.043</td>
<td>.088</td>
<td>.096</td>
<td>.050</td>
</tr>
<tr>
<td>12</td>
<td>.100</td>
<td>.157</td>
<td>.109</td>
<td>.127</td>
<td>.160</td>
<td>.125</td>
<td>.169</td>
<td>.177</td>
<td>.114</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.099</td>
<td>.073</td>
<td>.048</td>
<td>.139</td>
<td>.084</td>
<td>.110</td>
<td>.116</td>
<td>.072</td>
</tr>
<tr>
<td></td>
<td>.010</td>
<td>.034</td>
<td>.034</td>
<td>.008</td>
<td>.090</td>
<td>.040</td>
<td>.086</td>
<td>.090</td>
<td>.048</td>
</tr>
<tr>
<td>( m^*(c) )</td>
<td>.100</td>
<td>.148</td>
<td>.108</td>
<td>.119</td>
<td>.178</td>
<td>.120</td>
<td>.153</td>
<td>.176</td>
<td>.112</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.094</td>
<td>.072</td>
<td>.045</td>
<td>.154</td>
<td>.083</td>
<td>.100</td>
<td>.114</td>
<td>.070</td>
</tr>
<tr>
<td></td>
<td>.010</td>
<td>.031</td>
<td>.033</td>
<td>.007</td>
<td>.100</td>
<td>.033</td>
<td>.078</td>
<td>.089</td>
<td>.047</td>
</tr>
</tbody>
</table>

(a) Lag-length parameter, (b) Calculated with implied probabilities \( \bar{r}_{t} \) as in (6.11), (c) Optimally chosen lag-length parameter using Newey & West (1994). A dagger in each row indicates closest value to the nominal level.

The following table reports the empirical powers of the test statistics considered in the previous table with \( m \) chosen by Newey & West’s (1994) nonparametric method. The power of each test is calculated at 9 points in the interval \( \beta_{1} = [-.768, 0.068] \) using 1000 replications and Monte Carlo size corrected critical
6.2. BLOCKING AND SMOOTHING METHODS

values.

Empirical power for lognormal stochastic volatility model $n = 1000$

| $\beta$ | $-\beta$ | $-\beta$ | $-\beta$ | $-\beta$ | $-\beta$ | $-\beta$ | $-\beta$ | $-\beta$ | $-\beta$
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------
| $J$     | .504    | .302    | .148    | .104    | 0       | .054    | .072    | .124    | .154    |
| $D^{\text{DET}}$ | .574    | .376†   | .211    | .128    | 0       | .025    | .048    | .147    | .185    |
| $S^{\text{DET}}$ | .611†   | .358    | .220    | .134    | 0       | .049    | .062    | .094    | .134    |
| $L^{\text{ET}(a)}$ | .434    | .240    | .130    | .103    | 0       | .075†   | .085†   | .155†   | .175    |
| $D^{\text{CU}}$ | .562    | .364    | .234†   | .136†   | 0       | .024    | .038    | .062    | .105    |
| $L^{\text{CU}(a)}$ | .449    | .261    | .154    | .117    | 0       | .039    | .071    | .155†   | .188†   |
| $J^f$   | .367    | .214    | .111    | .073    | 0       | .004    | .054    | .094    | .121    |
| $D^{\text{DET},f}$ | .419    | .266    | .158    | .090    | 0       | .002    | .063†   | .121†   | .149†   |
| $S^{\text{DET},f}$ | .446†   | .254    | .165    | .096    | 0       | .042†   | .028    | .071    | .106    |
| $L^{\text{ET}(a),f}$ | .316    | .170    | .097    | .083    | 0       | .006    | .053    | .117    | .138    |
| $D^{\text{CU},f}$ | .410    | .258†   | .175†   | .123†   | 0       | .008    | .024    | .042    | .082    |
| $L^{\text{CU}(a),f}$ | .327    | .185    | .115    | .084    | 0       | .012    | .051    | .118    | .148    |

(a) Calculated with implied probabilities $\overline{\tau}_i$ as in (6.11)
Chapter 7

Other applications

In this chapter we focus on a number of recent applications of EL and related methods in the context of moment condition models that are not differentiable nor finite dimensional.

7.1 Non differentiable moment indicators

In this section we consider as in Parente & Smith (2005) GEL in the context of moment conditions models with non differentiable functions. In this case the difficulty is to show the asymptotic normality because the standard argument based on a mean value expansion cannot be used. As usual we assume that 

\[ E[g(\theta_0)] = 0 \]

Theorem 52 (Parente and Smith (2005)) Assume that (I) the parameter space \( \Theta \) is a compact set, (II) \( g(\theta) \) is continuous at each \( \theta \) a.s., (III) \( E[\sup_{\theta \in \Theta} \|g(\theta)\|^\alpha] < \infty \) for some \( \alpha > 2 \), (IV) \( \rho(v) \) is twice continuously differentiable in a neighbourhood of \( 0 \), (V) \( \Omega_0 \) is p.d., (VI) \( \theta_0 \in \text{int}\{\Theta\} \), (VII) \( E[g(\theta)] \) is differentiable at \( \theta_0 \) with derivative matrix \( G_0 = \partial E[g(\theta)] / \partial \theta' \) of full column rank, (VIII) for any \( \delta_n \to 0 \) \( \sup_{\|\theta - \theta_0\| \leq \delta_n} \|\hat{g}(\theta) - E[g(\theta)] - \hat{g}(\theta_0)\| = o_p(n^{-1/2}) \), (IX) \( a \) \( \sup_{\theta \in \Omega_0} \|G(\theta)\| < \infty \) and \( b \) \( G_0^{-1}G_0^tG_0 \) is nonsingular. Then

\[
\begin{align*}
\frac{n^{1/2}}{2} \left[ \hat{\lambda} \frac{1}{(\hat{\theta} - \theta_0)} \right] & \overset{d}{\to} N \left( \left[ \begin{array}{c} 0 \\ 0 \end{array} \right], \left[ \begin{array}{cc} \Psi_0 & 0 \\ 0 & (G_0^{-1}G_0^tG_0^{-1})^{-1} \end{array} \right] \right), \\
2n \left( \hat{P}_\rho(\hat{\theta}, \hat{\lambda}) - \hat{P}_\rho(0) \right) & \overset{d}{\to} \chi^2_{k-1} 
\end{align*}
\]

where \( \Psi_0 \) is an in Theorem 15.
Remark 53 Assumptions (I), and (III)-(VI) are as in Theorem 15. Assumption (II) allows the moment indicator to not be continuous on all of $\Theta$ for a given $z$. Assumption (VII) replaces differentiability of the moment indicator with differentiability of its expectation, whereas assumption (VIII) is a standard stochastic equicontinuity condition.

Proof. Consistency can be proved as in Theorem 15 and thus it will be assumed. The basic idea is to replace the (non differentiable) objective function $\hat{P}_p(\theta, \lambda)$ with a smooth well behaved (albeit infeasible) one, given by

$$\hat{P}^*_p(\theta, \lambda) = [-G_0(\theta - \theta_0)]' \lambda - \hat{g}(\theta_0)' \lambda - \lambda' \Omega_0 \lambda / 2$$

and define $[\hat{\lambda}', \hat{\theta}']'$ as $\min_{\lambda} \sup_{\lambda} \hat{P}^*_p(\theta, \lambda)$. The distribution of these estimators follows easily from the first order conditions using CLT and CMT. Following (Pakes & Pollard 1989) we show that

$$(1) \ n^{1/2} (\hat{\theta} - \theta_0) = O_p (1),$$

$$(2) (a) \ \hat{P}_p(\theta, \lambda) = \hat{P}^*_p(\theta, \lambda) + o_p (1) \ \text{uniformly},$$

$$(2) (b) \left\| \left[ \hat{\lambda}', \hat{\theta} \right]' - \left[ \hat{\lambda}', \theta \right]' \right\| = o_p \left( n^{-1/2} \right).$$

These conditions are sufficient to establish the asymptotic normality of $\hat{\theta} - \theta_0$ given the asymptotic normality of $\bar{\theta} - \theta_0$. To show (1) note that (VIII) implies that $\| E \left[ g(\bar{\theta}) \right] \| = O_p (n^{-1/2})$ while (VII) implies that $\| E \left[ g(\theta) \right] \| \geq C \| \theta - \theta_0 \|$ from which (1) follows evaluating the latter expression at $\hat{\theta}$. To show (2)(a) note that by Taylor expansion about 0, T, (VII) and (VIII)

$$\left| \hat{P}_p(\hat{\theta}, \hat{\lambda}) - [-G_0(\hat{\theta} - \theta_0)]' \lambda - \hat{g}(\theta_0)' \lambda - \hat{\lambda}' \Omega_0 \hat{\lambda} / 2 \right| \leq (7.1)$$

$$\left| \hat{\lambda} \left( \sum_{i=1}^n \rho_2 \left( \tilde{g}_i(\hat{\theta}) \right) g_i(\hat{\theta}) \hat{g}_i(\hat{\theta})' / n + \Omega_0 \right) \hat{\lambda} \right| \leq \| \hat{\lambda} \| \left( \| \hat{g}(\hat{\theta}) - \hat{g}(\theta_0) \| + \| G_0 \| \left( \| \hat{\theta} - \theta_0 \| \right) \right) + \| \hat{\lambda} \|^2 \times$$

$$\left\| \sum_{i=1}^n \rho_2 \left( \tilde{g}_i(\hat{\theta}) \right) g_i(\hat{\theta}) \hat{g}_i(\hat{\theta})' / n + \Omega_0 \right\| = o_p \left( n^{-1} \right).$$

By CLT $n^{1/2} (\hat{\theta} - \theta_0) = O_p (1)$; moreover by (VII) and (VIII) $\| E \left[ g(\hat{\theta}) \right] \| \leq \| G_0 \| \left( \| \hat{\theta} - \theta_0 \| \right) + o \left( \| \hat{\theta} - \theta_0 \| \right)$ and

$$\| \hat{g}(\hat{\theta}) \| \leq o_p \left( n^{-1/2} \right) + \| E \left[ g(\hat{\theta}) \right] \| = O_p \left( n^{-1/2} \right).$$
Thus as in (7.1) \( \left| \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) - \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) \right| = o_p (n^{-1}) \). Finally (2)(b) follows noting that by definition of \( \hat{\theta} \) and \( \hat{\lambda} \)

\[
\tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) - o_p (n^{-1}) \leq \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) \leq \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) + o_p (n^{-1})
\]

which implies that \( \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) = \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) + o_p (n^{-1}) \). Then as in Pakes & Pollard (1989) an expansion about \( \hat{\theta} \) using (IX) (a) gives

\[
\tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) - \tilde{P}_\rho (\hat{\theta}, \hat{\lambda}) = \left[ G_0 (\hat{\theta} - \tilde{\theta}) \right]' \hat{\lambda} = o_p (n^{-1})
\]

which implies that \( \left\| \hat{\theta} - \tilde{\theta} \right\| = o_p (n^{-1/2}) \) by VII. A similar arguments shows that \( \left\| \hat{\lambda} - \lambda \right\| = o_p (n^{-1/2}) \).

The proof of the second result is as in Theorem 7.

### 7.2 Conditional moment conditions models

In this section we consider how EL methods can be used in the context of (efficient) estimation and inference in conditional moment conditions. Let \((z_i, x_i)_{i=1}^n\) be i.i.d. on a data vector \((z, x)\) and assume that there exists a function \(g(z, \theta) : \mathbb{R}^l \rightarrow \mathbb{R}^k \) \((l \geq k)\)

\[
E [g(z, \theta_0) | x] = 0.
\]

(7.2)

By the law of conditional expectation (7.2) implies

\[
E [A(x, \theta_0) g(z, \theta_0)] = 0
\]

(7.3)

where \(A(\cdot)\) is a matrix of instrumental variables. An interesting question is to find an \(A(\cdot)\) that yields an asymptotically efficient estimator of \(\theta_0\). By the results of Chamberlain (1987) it is well-known that

\[
A(x, \theta_0) = E [G_0'|x] (E [g(z, \theta_0) g(z, \theta_0)' | x])^{-1}
\]

achieves the semiparametric efficiency bound. Newey (1993) showed that the asymptotic variance of the optimal GMM estimator based on \(g(z, \theta_0) \otimes q(x)\), where \(q(\cdot)\) is an approximating function corresponds to a minimum mean squared error of \(A(x, \theta_0) g(z, \theta_0)\) by linear combinations of \(g(z, \theta_0) \otimes q(x)\). Thus as the dimension of \(q(\cdot)\) grows with the sample size (at an appropriate rate) the variance of GMM approaches the semiparametric bound. There are many possible choices for the approximating function \(q(\cdot)\) including splines, power and Fourier series. Thus we can approximate (7.3) with

\[
E [g(z, \theta_0) \otimes q(x)] = 0
\]
and base EL estimation on this (infinite dimensional) set of moment conditions, that is \( \hat{\theta} = \max_{\theta \in \Theta} \sup_{\lambda} \left( 1 + \lambda' g_i(\theta) \otimes q(x) \right) \). More generally we can defined GEL estimator as \( \hat{\theta} = \min_{\theta \in \Theta} \sup_{\lambda} \hat{P}_{\rho}(g_i(\theta) \otimes q(x), \lambda) \).

**Theorem 54 (Donald, Imbens and Newey (2003))** Assume that (I) \( E \left[ q(x) q(x)' \right] \) is finite and for any \( A(x) \) such that \( E \| A(x) \|^2 < \infty E \left[ (A(x) - q(x)' \gamma)^2 \right] \to 0 \),

(II) there is a \( \zeta(m) \) such that for each \( m = m(n) \) there is a nonsingular matrix \( B \) such that \( \eta(x) = Bp(x) \) and sup \( \| q(x) \| \leq \zeta(m) \), \( \sigma_{\min}(E[B(x)\eta(x)B(x)']) > 0 \) and \( m^{1/2} \leq \zeta(m) \), (III) (i) \( E[q(z, \theta_0) | x] = 0 \) for a unique \( \theta_0 \), (ii) \( \Theta \) compact, (iii) \( E \left( \sup_{\theta \in \Theta} \| g(\theta) \|_2^\delta \right) < \infty \) for \( \delta > 2 \), (iv) for all \( \theta, \theta' \in \Theta \)

\[ \| g(\theta) - g(\theta') \| \leq \delta(z) \| \theta - \theta' \|_1^\alpha \] for \( \alpha > 0 \) and \( E[\delta(z)]^2 < \infty \), (v) \( \zeta(m) \) \( m^2/n \to 0 \). Then

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

where \( G_0 = E \left[ \partial q(\theta_0)/\partial \theta' | x \right] \) and \( \Omega_0 = E \left[ g(\theta_0) g(\theta_0)' | x \right] \).

**Proof.** Let \( \delta_n = o \left( n^{-1/2}/\zeta(m) \right) \) and \( \Lambda_n = \{ \lambda : \| \lambda \| \leq \delta_n \} \). Then

\[ \sup_{\theta \in \Theta} \max_{\lambda \in \Lambda_n} \| \lambda' g_i(\theta) \| \leq \delta_n \sup_{\theta \in \Theta} \| g_i(\theta) \| \zeta(m) = o_p(1) \]

and \( \Lambda_n \subseteq \hat{\Lambda}_n(\theta) \) w.p.a.1. Then as in the proof of Theorem 15 for any \( \bar{\theta} = \theta_0 + o_p(m) \) and \( \| \hat{g}(\bar{\theta}) \| = O_p \left( (m/n)^{1/2} \right) \) Taylor expansion \( \sup_{\lambda \in \hat{\Lambda}_n(\bar{\theta})} \hat{P}_{\rho}(\bar{\theta}, \lambda) \leq \rho_0 + O_p \left( m/n \right) \) and \( \| \hat{\lambda} \| \leq O_p \left( m/n^{1/2} \right) \). We next show that \( \| \hat{g}(\bar{\theta}) \| = O_p \left( m/n^{1/2} \right) \) where \( \bar{\theta} = \arg \min_{\theta \in \Theta} \hat{P}_{\rho}(\theta, \lambda) \) for \( \lambda \in \Lambda_n \). Let \( \bar{\lambda} = -\delta_n \hat{\lambda}(\bar{\theta}) / \| \hat{\lambda}(\bar{\theta}) \| \) and note that \( \hat{P}_{\rho}(\bar{\theta}, \bar{\lambda}) \geq \rho_0 + \delta_n \| \hat{g}(\bar{\theta}) \| - C \delta_n^2 \) and since \( m/(n\delta_n) = o \left( (m/n)^{1/2} \right) \)

\[ \| \hat{g}(\bar{\theta}) \| \leq O_p \left( m/n^{1/2} \right) \]. The consistency of \( \bar{\theta} \) follows by noting that

\[ \hat{P}_{\rho}(\bar{\theta}, \bar{\lambda}) = \hat{g}(\bar{\theta})' \hat{\Omega}(\bar{\theta})^{-1} \hat{g}(\bar{\theta}) \to 0, \]

continuity of \( E \left[ g(\theta)' \right] \Omega^{-1} E \left[ g(\theta)' \right] \) and ULLN for \( \hat{g}(\theta)' \hat{\Omega}(\theta)^{-1} \hat{g}(\theta) \). The latter two follows noting that

\[ \left| E \left[ g(\theta)' \right] \Omega^{-1} E \left[ g(\theta)' \right] - E \left[ g(\theta)' \right] \Omega^{-1} E \left[ g(\theta) \right] \right| \leq E[\delta(z)]^2 \| \bar{\theta} - \theta \|_{2}\alpha, \]

\( \hat{g}(\theta)' \hat{\Omega}(\theta)^{-1} \hat{g}(\theta) \) satisfies a LLN and

\[ \left| \hat{g}(\theta)' \hat{\Omega}(\theta)^{-1} \hat{g}(\theta) - \hat{g}(\theta)' \hat{\Omega}(\theta)^{-1} \hat{g}(\theta) \right| \leq \max_{l,k} \left| \hat{\Omega}(\theta)^{-1} \right|^2 \| \bar{\theta} - \theta \|_{2}\alpha = O_p(1) \]
7.2. CONDITIONAL MOMENT CONDITIONS MODELS

(cf. Lemma A5 of Donald, Imbens & Newey (2003)) and the ULLN follows by Newey (1991). The asymptotic normality follows by mean value expansion and CLT, noting that

\[
\left\| \sum_{i=1}^{n} \rho_2 \left( \hat{\lambda} g_i (\hat{\theta}) \right) g_i (\hat{\theta})' / n - \Omega_0 \right\| = O_p \left( \zeta (m) (m^2 / n)^{1/2} \right),
\]

\[
\left\| \sum_{i=1}^{n} \rho_1 \left( \hat{\lambda} g_i (\hat{\theta}) \right) G_i (\hat{\theta}) / n - G_0 \right\| = O_p \left( (m^2 / n)^{1/2} \right).
\]

**Remark 55** The rate conditions correspond to \( m^3 / n \to 0 \) for splines and \( m^4 / n \to 0 \) for power series. For GMM estimator the rate conditions are slightly weaker (i.e. \( m^2 / n \to 0 \) for splines and \( m^3 / n \to 0 \) for power series).

We now briefly consider consistent conditional moment tests. The GEL statistic has the usual form

\[
2n \left( \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) - \hat{P}_\rho (0) \right).
\]

If \( m \) was fixed then the asymptotic distribution would be \( \chi^2_{m-k} \). However \( m \to \infty \) so the statistic we consider is a standardised version, as suggested for example by DeJong & Bierens (1994).

**Theorem 56** Under the same assumptions of Theorem 54 with (v) strengthened to \( \zeta (m^2) m^3 / n \to 0 \) then

\[
2n \left( \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) - \hat{P}_\rho (0) \right) - (2 (lm - k)) / (2 (lm - k)) \overset{d}{\to} N (0, 1).
\]

**Proof.** By the usual arguments

\[
2n \left( \hat{P}_\rho (\hat{\theta}, \hat{\lambda}) - \hat{P}_\rho (0) \right) = n \hat{\gamma} (\hat{\theta})' \hat{\Omega} (\hat{\theta})^{-1} \hat{\gamma} (\hat{\theta}) + o_p (1)
\]

and the conclusion follows by noting that the following conditions (I) \( \left\| \hat{\Omega} (\hat{\theta}) - \Omega_0 \right\| = o_p \left( (lm)^{-1/2} \right) \), (II) \( \sigma_{\min} (\Omega) > 0 \), (III) \( \left\| \hat{G} (\hat{\theta}) - G_0 \right\| \overset{p}{\to} 0 \), (IV) \( G_0^{1/2} G_0^{-1} G_0 \) is bounded, (V) \( E \left[ \hat{g} (\theta_0)' \Omega_0 \hat{g} (\theta_0) \right] / (lm n^{1/2}) \to 0 \) are satisfied. These conditions imply that

\[
n \left( \hat{\gamma} (\hat{\theta})' \hat{\Omega} (\hat{\theta})^{-1} \hat{\gamma} (\hat{\theta}) - n \hat{\gamma} (\theta_0)' \Omega_0^{-1} \hat{g} (\theta_0) \right) / (2 (lm - k)) \overset{p}{\to} 0
\]

and that

\[
\left( n \hat{\gamma} (\theta_0)' \Omega_0^{-1} \hat{g} (\theta_0) - (lm - k) / (2 (lm - k)) \right) \overset{d}{\to} N (0, 1)
\]
A alternative approach to obtain an efficient estimator of \( \theta_0 \) in (7.2) is to use the so-called conditional or local EL proposed by Kitamura, Tripathi & Ahn (2004). This method avoids the explicit estimation of the matrix \( A(x, \theta_0) \) of optimal instruments by using a local (i.e. kernel based) EL to estimate the conditional moment restriction itself.

Let \( w_{ij} = K_h(x_i - x_j) / \sum_{j=1}^{n} K_h(x_i - x_j) \) denote a set of positive weights used to carry out the localisation, where \( K_h(\cdot) \) is a kernel function, and let \( \pi_{ij} = \Pr(z = z_j | x = x_i) \) denote the conditional probability supported on the observed sample \((x_i)_{i=1}^{n} \times (z_j)_{j=1}^{n}\). Then as in (2.2) we can define

\[
\max_{\pi_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \log \pi_{ij} \quad \text{s.t.} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_{ij} = 1 \quad \text{and} \quad \sum_{j=1}^{n} \pi_{ij} g(z_j, \theta_0) = 0,
\]

whose solution is

\[
\hat{\pi}_{ij} = w_{ij} / \left( 1 + \hat{\lambda}_i g(z_j, \theta_0) \right) \quad i = 1, ..., n
\]

where the Lagrange multipliers \( \hat{\lambda}_i \) satisfy

\[
\sum_{j=1}^{n} \pi_{ij} g(z_j, \theta_0) / \left( 1 + \hat{\lambda}_i g(z_j, \theta_0) \right) = 0.
\]

We can then define the maximum conditional EL (CEL) estimator for \( \theta_0 \) as

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} CEL(\theta)
\]

where

\[
CEL(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,n} w_{ij} \log \left( w_{ij} / \left( 1 + \hat{\lambda}_i g(z_j, \theta) \right) \right),
\]

\( T_{i,n} = I \left\{ \sum_{j=1}^{n} K_h(x_i - x_j) / (nh) \geq h^\tau \right\} \) with \( \tau \in (0, 1) \) is a sequence of trimming functions included to avoid the well-known denominator problem of kernel estimators.

**Theorem 57 (Kitamura, Tripathi and Ahn (2004))** Assume that (I) \( E[g(z, \theta) | x] \neq 0 \) for each \( \theta \neq \theta_0 \) for every \( x \in \chi_\theta \) such that \( \Pr(x \in \chi_\theta) \), (II) \( E[\sup_{\theta \in \Theta} \|g(z, \theta)\|_\alpha] < \infty \) \( \alpha \geq 8 \), (III) \( K \) is a symmetric continuously differentiable pdf with support \([1, 1]\) and bounded away from 0 on \([-\alpha, \alpha]\) for some \( \alpha \in (0, 1) \), (IV) \( 0 < l(x) \leq \sup_{x \in \mathbb{R}} l(x) < \infty \), \( l(x) \) is twice continuously differentiable with \( \sup_x \|\partial^2 l(x) / \partial x \partial x'\| < \infty \), where \( l(x) \) is the Lebesgue density of \( x \), (V) \( E\|x\|^{1+\beta} < \infty \) for some \( \beta > 0 \), (VI) \( E[\sup_{\theta \in \Theta} \|\partial g(z, \theta) / \partial \theta'\|] < \infty \), (VII) \( \sup_{\theta \in \Theta} \|...
Proof. See Kitamura et al. (2004) □

We now briefly consider testing restrictions on \( \theta_0 \). As in Theorem 10 we want to test \( H_0 : h(\theta_0) = 0 \). Let

\[
CELR = 2 \left( CEL \left( \tilde{\theta} \right) - CEL \left( \hat{\theta} \right) \right)
\]

denote the conditional empirical likelihood ratio where \( \hat{\theta} \) is the constrained estimator.

**Theorem 58** Under the same assumptions of Theorem 28, and \( H_0 \)

\[
CELR \overset{d}{\to} \chi^2_q.
\]

**Proof.** Note that the results of Theorem 28 applies to the modified parameter space \( \Theta_0 = \{ \theta : h(\theta_0) = 0 \} \) which is still compact. Thus \( \tilde{\theta} \) is also consistent. Then by a second order Taylor expansion about \( \hat{\theta} \) gives

\[
CELR = n \left( \hat{\theta} - \theta_0 \right)' \left[ \partial^2 CEL \left( \hat{\theta} \right) / \partial \theta \partial \theta' / n \right] \left( \hat{\theta} - \theta_0 \right).
\]

By results of Kitamura et al. (2004) \( \partial^2 CEL \left( \hat{\theta} \right) / \partial \theta \partial \theta' / n \overset{p}{\to} G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \) and

\[
n^{1/2} \left( \hat{\theta} - \theta_0 \right) = \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} G_{0x}^{-1} \Omega_{0x}^{-1} n^{1/2} \tilde{\theta} / \theta_0 + o_p(1).
\]

Moreover by standard manipulations \( n^{1/2} \left( \tilde{\theta} - \hat{\theta} \right) = \Upsilon(\theta_0) G_{0x}^{-1} \Omega_{0x}^{-1} n^{1/2} \tilde{\theta} / \theta_0 + o_p(1) \) where

\[
\Upsilon(\theta_0) = \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} \left[ I - H_0' \left( H_0 \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} H_0' \right)^{-1} H_0 \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} \right]
\]

so that

\[
n^{1/2} \left( \hat{\theta} - \theta_0 \right) = \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} H_0' \left( H_0 \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} H_0' \right)^{-1} H_0 \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} \times \left( G_{0x}^{-1} \Omega_{0x}^{-1} G_0 \right)^{-1} \tilde{\theta} / \theta_0.
\]
Let $\xi(\theta_0) = G'_{0x}\Omega^{-1}_{0x} n^{1/2} \hat{\theta}_0$ and note that $\xi(\theta_0) \overset{d}{\rightarrow} N(0, G'_{0x}\Omega^{-1}_{0x} G_{0x})$. Then (7.5) is

$$CELR = \xi(\theta_0)' \left( G'_{0x}\Omega^{-1}_{0x} G_{0x} \right)^{-1} H_0' \left( H_0 \left( G'_{0x}\Omega^{-1}_{0x} G_{0x} \right)^{-1} H_0' \right)^{-1} \xi(\theta_0)$$

and since $\left( G'_{0x}\Omega^{-1}_{0x} G_{0x} \right)^{-1} H_0' \left( H_0 \left( G'_{0x}\Omega^{-1}_{0x} G_{0x} \right)^{-1} H_0' \right)^{-1} H_0$ is idempotent with rank $q$ the conclusion follows by standard results.

Smith (2007) extends the CEL to ECR statistics, that is

$$\max_{\pi_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,n} w_{ij} \left[ (\pi_{ij}/w_{ij})^{-\gamma} - 1 \right] / \left[ \gamma (\gamma + 1) \right] \text{ s.t.}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1 \text{ and } \sum_{j=1}^{n} \pi_{ij} g(z_j, \theta) = 0,$$

and shows that all conditional ECR estimators defined as

$$\hat{\theta} = \arg \max_{\theta} CECR(\theta)$$

where

$$CECR(\theta) = \frac{2}{\gamma (\gamma + 1)} \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,n} \left[ 1 + \hat{\zeta}_i + \hat{\lambda}_i g(z_j, \theta) \right]^{\gamma/(\gamma + 1)} - 1$$

are asymptotically equivalent to the conditional EL estimator.

Zhang & Gijbels (2003) also consider CEL\(^1\), however their interest is explicitly in parametric and nonparametric (conditional) regressions models as opposed to the conditional moment restrictions themselves as in Kitamura et al. (2004) and Smith (2007). Here we consider some of their results for nonparametric regression models. We first give an example to illustrate and motivate their approach.

**Example 59** Consider a nonparametric regression model $y_i = \theta_0(x_i) + \varepsilon_i$ where the unobservable errors $\varepsilon_i$ satisfy $E[\varepsilon_i^2 | x_i] = \theta(x_i)^2 + 1$ a.s. and $E[\varepsilon_i^{2(k-1)+1} | x_i] = 0$ a.s. for $1 \leq k \leq k_0 - 1$. Then the $\mathbb{R}^{k_0}$-valued vector

$$g(z_i, \theta) = (y_i - \theta(x_i))^2 - \left( \theta(x_i)^2 + 1 \right), (y_i - \theta(x_i))^{2(k-1)+1}$$

satisfies the conditional moment condition $E[g(z_i, \theta) | x_i] = 0$ a.s.

\(^1\)Note that they call their procedure sieve empirical likelihood (SEL).
Zhang & Gijbels (2003) suggest to use CEL to simultaneously estimate the unknown regression function \( \theta(\cdot) \) and the unknown conditional distribution of the observations. We assume that \( \theta \) belongs to a smooth class \( \Theta \) and is equipped with norm \( \|\theta\|^2 = E \left[ \|\theta(x_i)\|^2 | x_i \right] \). Let \( z_i = [y_i, x_i'] \in \mathbb{R}^{k+1} \),

\[
F_g = \{ g(z_i, \theta) \mid \|\theta - \theta_0\| \leq \delta_n \},
\]

\[
F_{gg'} = \{ g(z_i, \theta) g(z_i, \theta') \mid \|\theta - \theta_0\| \leq \delta_n \}
\]

for any positive \( \delta_n \to 0 \), and \( \theta(\theta^*, s) \in \Theta \) (0 \( s \leq 1 \)) denote a path\(^2\) from \( \theta \) to \( \theta^* \). Also define

\[
\partial g(\theta(\cdot)) / \partial \theta = \lim_{s \to 0} \left( g(\theta(\theta^*, s)) - g(\theta) \right) / s
\]

\[
\|\partial g(\theta(\cdot)) / \partial \theta\| = \sup_{\theta^* \in \Theta, \theta^* \neq \theta} \| \partial g(\theta(\cdot)) / \partial \theta \|_{\theta^* - \theta} / \| \theta^* - \theta \|
\]

and for a given class of functions \( \mathcal{F} \) let

\[
H^B(\epsilon, L_2(P), \mathcal{F}) = \log N^B(\epsilon, L_2(P), \mathcal{F})
\]

denote the \( \epsilon \) entropy with bracketing, where \( N^B(\cdot) \) is the bracketing covering number\(^3\).

The following theorem establishes consistency and the convergence rate of CEL defined as

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} CEL(\theta) = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \log \left( w_{ij} / \left( 1 + \tilde{\lambda}_i g(z_j, \theta) \right) \right).
\]

**Theorem 60 (Zhang and Gijbels (2003))** Assume that (I) \( \sup_{\|\theta - \theta_0\| \geq \delta} \| E[g(z_i, \theta)] \| > 0 \) a.s., (II) \( \int K(\|s\|) ds = 1 \), \( \int K(\|s\|) s ds = 0 \), \( \int K(\|s\|)^{(1+\delta)/\delta} ds < \infty \) for any \( \delta > 0 \), \( |K(s_1) - K(s_2)| \leq C |s_1 - s_2| \) for any \( s_j \in \mathbb{R} \) \( j = 1, 2 \), the bandwidth \( h \) satisfies \( c_0 \leq h^k n^\alpha \leq c_1(III) \) \( x_i \) has compact and convex support \( X_0 \), (IV) \( E \sup_{\|\theta - \theta_0\| \leq \delta, \theta \in \Theta} E[g_j^2(z_i, \theta)] < \infty \) \( j = 1, \ldots, k \) \( \alpha > 4 \), (V) \( \sup_{\|\theta - \theta_0\| \leq \delta, \theta \in \Theta} \sup_{x_i \in X_0} \| \partial E[g_j(z_i, \theta)] / \partial \theta \| < \infty \) \( j = 1, \ldots, k \), (VI)

\[
\sup_{\|\theta - \theta_0\| \leq \delta, \theta \in \Theta} \sup_{x_i \in X_0} \| \partial E[g_j(z_i, \theta)] / \partial \theta \| < \infty \) \( j = 1, \ldots, k \), (VII)

\[
\sup_{\|\theta - \theta_0\| \leq \delta, \theta \in \Theta} \sup_{x_i \in X_0} \left| E[g_j(z_i, \theta(x_i + u)) | x_i] - E[g_j(z_i, \theta(x_i)) | x_i] \right| \to 0
\]

\[
\sup_{\|\theta - \theta_0\| \leq \delta, \theta \in \Theta} \sup_{x_i \in X_0} \left| E[g_{jl}(z_i, \theta(x_i + u)) | x_i] - E[g_{jl}(z_i, \theta(x_i)) | x_i] \right| \to 0
\]

\(^2\)For example \( \theta(\theta^*, s) = (1 - s) \theta + s \theta^* \).

\(^3\)The bracketing covering number is the smallest number \( m \) such that for any \( f \in \mathcal{F} \) there exist \( f_k^u \leq f_k^b \) with \( \max_{1 \leq k \leq m} \int |f_k^u - f_k^b|^2 dP \leq \epsilon \) such that \( f_k^u \leq f \leq f_k^b \).
as \( u \to 0 \) where \( g_{jl} (\cdot) = g_j (\cdot) g_l (\cdot) \) \((j,l = 1, \ldots, k)\), (VIII)

\[
\sup_{\|x-x_0\| \leq \delta, \theta \in \Theta} \sigma_{\min} \left( E \left[ g (z_i, \theta) g (z_i, \theta)' \middle| x_i \right] \right)
\]

is positive definite (IX)

\[
H^B (\epsilon, L_2 (P), F_g) \leq A (P) (\epsilon / \delta_0)^{-\beta} \\
H^B (\epsilon, L_2 (P), F_{g''}) \leq B (P) (\epsilon / \delta_0)^{-\beta},
\]

(X) there exists a measurable function \( M (z) \) and a \( \zeta > 0 \) such that for any \( \theta_j \in \Theta \) \((j = 1, 2)\)

\[
|g_j (z_i, \theta_1) - g_j (z_i, \theta_2)| \leq M (z) (\theta_1 - \theta_2) \\
\sup_{x_i \in X_0} E \left[ \exp (\zeta M (z)) |x_i| \right] < \infty \ a.s., \\
|\partial E [g_j (z_i, \theta_1)] |x_i|/\partial \theta - \partial E [g_j (z_i, \theta_2)] |x_i|/\partial \theta| \leq C |\theta_1 - \theta_2|
\]

(XI)

\[
\sup_{x_i \in X_0} \left\| E \left[ g (z_i, \theta_0 (x_i + u)) g (z_i, \theta_0 (x_i + u))' \right] - E \left[ g (z_i, \theta_0 (x_i)) g (z_i, \theta_0 (x_i))' \right] \right\| \leq C |u|
\]

If (I)-(IX) hold and \( 0 < \eta < 2 (\alpha - 4) / (\alpha (2 + \beta)) \) then \( \hat{\theta} \overset{p}{\rightarrow} \theta_0 \). In addition if (X-XI) hold then for any \( 0 \leq \pi \leq \pi^* \) and \( \eta_1 (\pi) < \eta < \eta_2 (\pi) \)

\[
\left\| \hat{\theta} - \theta_0 \right\| = O_p \left( \max \left\{ n^{-1/(2+\beta^*)}, n^{-\pi/(1-\beta^*)} \right\} \right)
\]

where \( \beta^* = 1/(q+\tau) \), \( \zeta_1 = 2/(2+\beta^*) \), \( \zeta_2 = 2(1-\beta^*)/(2-\beta^*) \), \( 1 \leq \zeta \leq 2 \),

\[
\pi^* = \min \left\{ \frac{1}{3 \left( \frac{2}{1-\beta^*} - \zeta_1 - \zeta_2 \right) + 2 \zeta_2 - \zeta_1}, \frac{2}{(6+\beta^*) \left( \frac{2}{1-\beta^*} - \zeta_2 - 1 \right) + 4 \zeta_2 - 2 \beta^* - \zeta_1 (2-\beta^*)} \right\} \\
\frac{1}{3 \left( \zeta_1 - \zeta_2 \right) + 2 \zeta_2 - \zeta_1}, \frac{2}{(6+\beta^*) \left( \zeta_1 (2-\beta^*) + 4 \zeta_2 - 2 \beta^* - \zeta_1 (2-\beta^*) \right)}
\]

\[
\eta_1 (\pi) = \min \left\{ \pi (\zeta_1 - \zeta_2), \frac{2\pi}{1-\beta^*} - \pi (1 + \zeta_2) \right\} \\
\eta_2 (\pi) = \min \left\{ (1 - 2\zeta_2 \pi + \zeta_1 \pi) / 3, 2 (1 - 2\zeta_2 \pi + \beta_2 \pi + \zeta_1 (2 - \beta^*) \pi) / (6 + \beta^*) \right\}
\]

**Proof.** Under the assumptions it is possible to show that

\[
\sup_{i} \left( 1 + \max_{i} \left\| g (z_i, \theta) \right\|_1 \right) \left\| \sum_{j=1}^{n} w_{ij} g (z_i, \theta) \right\| = o_p (1)
\]
uniformly in $x_i \in X_0, \theta \in \Theta$ and $\|\theta - \theta_0\| \leq \delta_n < n^{-1/\alpha}$, which implies by standard arguments (cf. proof of Theorem 4) that

$$\hat{\lambda}_i = \left( \sum_{j=1}^n w_{ij} g(z_j, \theta) g(z_j, \theta)' \right)^{-1} \sum_{j=1}^n w_{ij} g(z_j, \theta) + o_p(1)$$

uniformly in $x_i \in X_0, \theta \in \Theta$ and $\|\theta - \theta_0\| \leq \delta_n < n^{-1/\alpha}$. Next use ULLN show that (truncated versions) of $\sum_{i=1}^n \sum_{j=1}^n w_{ij} g(z_j, \theta) / n$ and $\sum_{i=1}^n \sum_{j=1}^n w_{ij} g(z_j, \theta) g(z_j, \theta)' / n$ converge uniformly to their expected values and proceed (similarly to the proof of Theorem 7) to show that

$$\Pr \left( \sup_{\|\theta - \theta_0\| \geq \delta} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log \left( 1 + \lambda_i g(z_j, \theta) \right) - \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log \left( 1 + \lambda_i g(z_j, \theta_0) \right) \right) \rightarrow 1.$$

To obtain the convergence rate note that

$$2CEL(\theta) = - \sum_{i=1}^n \left( \sum_{j=1}^n w_{ij} g(z_j, \theta)' \right) \left( \sum_{j=1}^n w_{ij} g(z_j, \theta) g(z_j, \theta)' \right)^{-1} \left( \sum_{j=1}^n w_{ij} g(z_j, \theta) \right) + o_p(1)$$

and that as $\|\hat{\theta} - \theta_0\| \leq \delta_n \rightarrow 0$

$$2(CEL(\theta) - CEL(\theta_0)) = - \sum_{i=1}^n \left( \sum_{j=1}^n w_{ij} g(z_j, \theta)' \right) \left( E [g(z_i, \theta) g(z_i, \theta)' | x_i] \right)^{-1} \left( \sum_{j=1}^n w_{ij} g(z_j, \theta) \right) + o_p(1).$$

The convergence rate is then obtained using the same approach used by Shen & Wong (1994), that consists of improving iteratively the rate by obtaining increasingly faster uniform approximation rates for $CEL(\theta)$ - see Zhang & Gijbels (2003) for more details.

**Remark 61** When $\beta^* \leq 1/4, \|\hat{\theta} - \theta_0\| = O_p \left( n^{-1/(2+\beta^*)} \right)$, which is the optimal convergence rate in ordinary nonparametric regression.

### 7.2.1 Applications of Conditional EL

In this section we briefly discuss two recent applications of the CEL approach.
CHAPTER 7. OTHER APPLICATIONS

Specification testing

Tripathi & Kitamura (2003) use CEL to test whether (7.2) holds over a compact set $X_0$ that is $E[g(z, \theta_0)|x] = 0$ a.s. for $\theta_0 = \Theta$ and $x \in X_0$. The same Lagrangian argument of the previous section shows that the CEL test statistic is

$$2CEL(\hat{\theta}) = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} I_i w_{ij} \log \left( w_{ij} / \left( 1 + \hat{\lambda}_i g \left( z_j, \hat{\theta} \right) \right) \right)$$

where $I_i = I \{ x_i \in X_0 \}$. Tripathi & Kitamura (2003) show that

$$2CEL(\hat{\theta}) - h^{-s} q c_1(K) / \left( 2h^{-s} q \right)^{1/2} c_2(K) \xrightarrow{d} N(0, 1)$$

where $c_1(K) = \int K(u)^2 \, du$ and $c_2(K) = \int \left( \int K(v) K(u-v) \, dv \right)^2 \, du$.

It is interesting to note that compared to other commonly used procedures $CEL(\hat{\theta})$ has two advantages. Assume that $g$ and $x$ are scalar and consider a weighted sum of squared residuals of a kernel regression (Ait-Sahalia, Bickel & Stoker 2001)

$$S_a = h \sum_{i=1}^{n} \sum_{j=1}^{n} a(x_i) \left[ w_{ij} g \left( z_j, \hat{\theta} \right) \right]^2$$

where $a(\cdot) : [0, 1] \to \mathbb{R}^+$ is a function such that $\int a(x) \, dx = 1$. An example of $S_a$ is the statistic proposed by Hardle & Mammen (1993). Let $\sigma^2(x) = V(g(z_j, \theta_0) | x)$ denote the conditional variance and note that

$$T_{s_a} = h^{-1/2} \left( S_a - c_1(K) \int \sigma a \, dx \right) / \left( 2 c_2(K) \int \sigma^2 a^2 \, dx \right)^{1/2} \xrightarrow{d} N(0, 1)$$

First note that, as opposed to $CEL$, $T_{s_a}$ is not invariant to reparameterisations of $g$. Second $CEL$ has an advantage in terms of power. To be specific let $E[g(z, \theta_n) | x] = n^{-1/2} h^{s/4} \delta(x)$ a.s.

denote a sequence of alternative hypothesis, let $f(x)$ denote the distribution of $x$ and

$$\mu(a, \delta) = \int_0^1 \delta^2(x) f(x) a^2(x) / \left( 2 c_2(K) \int \sigma^2 a^2 \, dx \right)^{1/2}$$

Then $T_{s_a} \xrightarrow{d} N(\mu(a, \delta), 1)$ and thus the asymptotic local power of the test with critical value $c_\alpha$ is

$$\pi(a, \delta) = 1 - \Phi(c_\alpha - \mu(a, \delta))$$

Given that $\delta$ is unknown one way to evaluate $\pi(a, \delta)$ is to use the notion of average local power, i.e. $\bar{\pi}(a, \delta) = \int \pi(a, \delta) \, dP_\delta$ where $P_\delta$ is the distribution
under the alternative. In case of parametric likelihood functions the latter is typically known, however in the present context of moment conditions models this is clearly not the case. Tripathi & Kitamura (2003) solve this problem by using the nonparametrically estimated distribution of the \( x_i \) that is consistent with the conditional moment condition

\[
\tilde{\delta}(x) = \sum_{i=1}^{n} K_h(x_i - x) g(x_i, \theta) / \sum_{j=1}^{n} K_h(x_i - x)
\]

They then redefine the average local power as

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

where \( C[0,1] \) is the set of continuous function over \([0,1]\) and show using some calculus of variations that \( \pi(a, \tilde{\delta}(x)) \) is maximised at

\[
a = 1/ \left( \sigma(x) \int_{0}^{1} \sigma^{-2}(x) dx \right)
\]

which happens to be the value of \( a \) that the statistic \( T_{sa} \) takes if we are considering the CEL. Thus CEL has maximal average local power.

Smith (2007) shows that the CCR statistic

\[
CCR(\theta) = \frac{2}{\gamma(\gamma + 1)} \sum_{i=1}^{n} \sum_{j=1}^{n} I_i \left[ (1 + \hat{\zeta}_i + \hat{\lambda}_i g(z_j, \hat{\theta}))^{\gamma/(\gamma+1)} - 1 \right]
\]

can be used to test (7.2). To be specific he shows that under \( H_0 \)

\[
(CCR(\theta) - h^{-s} q_1(K)) / \left( (2h^{-s} q)^{1/2} c_2(K) \right) \xrightarrow{d} N(0,1).
\]

By the asymptotic equivalence between \( CEL \) and \( CECR \) it is expected that all of the members of the latter enjoy the optimality property of the former in terms of average local power.

Non nested conditional moment restrictions

Otsu & Whang (2007) use CEL to develop three non nested tests: moment encompassing, Cox-type and efficient score encompassing test. Consider two competing conditional moment restriction models

\[
H_0^g : E[g(z_i, \theta) | x_i] = 0 \text{ if } \theta = \theta_0
\]
\[
H_0^h : E[h(z_i, \beta) | x_i] = 0 \text{ if } \beta = \beta_0
\]
Let $CEL_g(\theta)$ denote the conditional EL as defined in Theorem 57, let $\hat{\theta}$ denote the Conditional MELE and similarly for model $h$. If the model is misspecified $\hat{\theta}$ and $\hat{\beta}$ converge to their pseudo-true value defined as

$$\hat{\theta}^* = \min_{\theta \in \Theta} E \left[ E \sup_{\lambda} \left( \log (1 + \lambda' g(z_i, \theta)) | x_i \right) \right],$$

$$\hat{\beta}^* = \min_{\beta \in B} E \left[ E \sup_{\lambda} \left( \log (1 + \lambda' h(z_i, \beta)) | x_i \right) \right].$$

To compute the three non nested statistics Otsu & Whang (2007) use the implied probabilities

$$\hat{\pi}_{ij}^g = \frac{w_{ij}}{1 + \lambda_i g(z_j, \hat{\theta})}, \quad \hat{\pi}_{ij}^h = \frac{w_{ij}}{1 + \lambda_i h(z_j, \hat{\beta})} \quad (7.6)$$

where $w_{ij}$ are the kernel weights $\hat{\theta}$ and $\hat{\beta}$ are any $n^{1/2}$ consistent estimators. Let $m_{ij}(\theta, \beta) = m_{ij}(z_j, x_i, \theta, \beta)$ denote a moment indicator that can be expressed as $M(x_i, \theta, \beta) \tilde{m}(z_j, \theta, \beta)$ for an appropriate matrix $M(\cdot)$ and vector $\tilde{m}(\cdot)$. Typically $M(\cdot) = I$ and $\tilde{m}(\cdot) = h(z_j, \beta)$ (i.e. it is the alternative model). Consider now the difference between the moment indicators evaluated at the implied probabilities (7.6) and those evaluated at the unconstrained one $w_{ij}$

$$T_m(\hat{\theta}, \hat{\beta}) = \sum_{i=1}^n \sum_{j=1}^n \hat{\pi}_{ij}^g m_{ij}(\hat{\theta}, \hat{\beta}) / n - \sum_{i=1}^n I_i \sum_{j=1}^n w_{ij} m_{ij}(\hat{\theta}, \hat{\beta}) / n$$

where as before $I_i = I \{ x_i \in X_0 \}$ is trimming function for a subset $X_0$ of $X$. Under the null hypothesis that $H_0$ is correct $T_m \overset{p}{\rightarrow} 0$ otherwise $T_m$ diverges. Then we can define the CEL-based moment encompassing test statistic for $H_0^g$ as

$$T_m^2 = nT_m(\hat{\theta}, \hat{\beta})' \phi_m(\hat{\theta}, \hat{\beta})^{-g} T_m(\hat{\theta}, \hat{\beta})$$

where "$^{-g}$" denotes generalised inverse, and

$$\phi_m(\theta, \beta) = \sum_{i=1}^n \psi_{im}(\theta, \beta) \psi_{im}(\theta, \beta)' / n,$$

$$\psi_{im}(\theta, \beta) = -I_i M(x_i, \theta, \beta) \hat{J}_i(\theta, \beta) \hat{\Omega}_i^g(\theta)^{-1} g(z_i, \theta) + \hat{H}_m(\theta, \beta) \Delta \psi_i(\theta)$$

$$\hat{J}_i(\theta, \beta) = \sum_{j=1}^n w_{ij} g(z_j, \theta) \tilde{m}_{ij}(z_j, \theta, \beta)' / n, \quad \hat{\Omega}_i^g(\theta) = \sum_{j=1}^n w_{ij} g(z_j, \theta, \beta)' g(z_j, \theta) / n,$$

$$\hat{H}_m(\theta, \beta) = \sum_{i=1}^n I_i M(x_i, \theta, \beta) \hat{J}_i(\theta, \beta) \hat{\Omega}_i^g(\theta)^{-1} \hat{G}_i(\theta),$$

$$\hat{G}_i(\theta) = \sum_{j=1}^n w_{ij} (\partial g(z_j, \theta) / \partial \theta)' / n,$$
7.2. CONDITIONAL MOMENT CONDITIONS MODELS

and

$$n^{1/2} (\hat{\theta} - \theta_0) = -\Delta^{-1} \sum_{j=1}^n \psi_i (\theta_0) + o_p (1). \quad (7.7)$$

To define a Cox-type conditional non-nested statistic consider the difference between the quadratic forms

$$T_c (\hat{\theta}, \hat{\beta}) = \sum_{i=1}^n I_i \hat{h}_i^q (\hat{\beta})' \hat{\Omega}_i^h (\hat{\beta})^{-1} \hat{h}_i^q (\hat{\beta}) / n - \sum_{i=1}^n I_i \hat{h}_i (\hat{\beta})' \hat{\Omega}_i^h (\hat{\beta})^{-1} \hat{h}_i (\hat{\beta}) / n$$

where

$$\hat{h}_i^q (\beta) = \sum_{j=1}^n \hat{g}_{ij}^q h (z_j, \beta), \quad \hat{h}_i (\beta) = \sum_{j=1}^n w_{ij} h (z_j, \beta), \quad \hat{\Omega}_i^h (\beta) = \sum_{j=1}^n w_{ij} h (z_j, \beta) h (z_j, \beta)' / n.$$

The CEL-base Cox test statistic is

$$C_g = n^{1/2} T_c (\hat{\theta}, \hat{\beta}) / \hat{\phi}_c^{1/2}$$

where

$$\hat{\phi}_c = \sum_{i=1}^n \psi_{ic} (\hat{\theta}, \hat{\beta})^2 / n,$$

$$\psi_{ic} (\theta, \beta) = -2 I_i \hat{h}_i (\beta) \hat{\Omega}_i^h (\theta)^{-1} \hat{h}_i (\beta) \hat{\Omega}_i^q (\theta)^{-1} \hat{g} (z_i, \theta) + \hat{H}_c (\theta, \beta) \Delta \psi_i (\theta)$$

$$\hat{J}_i^h (\theta, \beta) = \sum_{j=1}^n I_i \hat{g}_{ij} (z_j, \beta) h (z_j, \beta)' / n,$$

$$\hat{H}_c (\theta, \beta) = 2 \sum_{j=1}^n I_i \hat{h}_i (\beta) \hat{\Omega}_i^h (\theta)^{-1} \hat{J}_i^h (\theta, \beta) \hat{G}_i (\theta).$$

Finally the CEL-based efficient score encompassing test, which focuses on the probability limit of the asymptotic linear form of the asymptotically efficient estimators for \( \beta_0 \) in \( H_0^h \), that is

$$n^{1/2} (\hat{\beta} - \beta_0) = -n^{-1/2} I^h (\beta_0) \sum_{i=1}^n G_i (\beta)' \Omega_i^h (\beta)^{-1} h (z_i, \beta)$$

where

$$I^h (\beta) = E \left[ G_i^h (\beta)' \Omega_i^h (\beta)^{-1} G_i^h (\beta) \right], \quad G_i^h (\beta) = E \left[ \partial h (z_i, \beta) / \partial \beta' | x_i \right],$$

$$\Omega_i^h (\beta) = E \left[ h (z_i, \beta) h (z_i, \beta)' | x_i \right].$$
Let \( \hat{G}_i^b (\beta) = \sum_{j=1}^{n} w_{ij} \partial h (z_j, \beta) / \partial \beta' \); similarly to the moment encompassing test consider the difference between two efficient scores

\[
T_s (\hat{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n} I_i \hat{G}_i^b (\hat{\beta}) \hat{\Omega}_i^b (\hat{\beta})^{-1} \hat{\pi}_i^b \hat{h}_{ij} (\hat{\beta}) / n - \sum_{i=1}^{n} \sum_{j=1}^{n} I_i \hat{G}_i^b (\hat{\beta}) \hat{\Omega}_i^b (\hat{\beta})^{-1} w_{ij} \hat{h}_{ij} (\hat{\beta}) / n
\]

can be used to test \( H_{0}^s \). The CEL based efficient score encompassing test is

\[
T_s^2 = n T_s (\hat{\beta})' \Phi_s (\hat{\beta})^{-g} T_s (\hat{\beta})
\]

where

\[
\Phi_s (\beta) = \sum_{i=1}^{n} \psi_{is} (\beta) \psi_{is} (\beta)' / n,
\]

\[
\psi_{is} (\theta, \beta) = - I_i \hat{G}_i^b (\beta) \hat{J}_i^b (\theta, \beta) \hat{\Omega}_i^g (\theta)^{-1} g (z_i, \theta) + \hat{H}_s (\beta) \Delta \psi_i (\theta)
\]

\[
\hat{J}_i^b (\theta, \beta) = \sum_{j=1}^{n} w_{ij} g (z_j, \theta) h (z_j, \beta)' / n,
\]

\[
\hat{H}_c (\theta, \beta) = 2 \sum_{j=1}^{n} I_i \hat{h}_i (\beta) \hat{\Omega}_i^b (\beta)^{-1} \hat{J}_i^b (\theta, \beta) \hat{\Omega}_i^g (\theta)^{-1} \hat{G}_i (\theta) / n.
\]

Under regularity conditions that are similar to those of Theorem 57 Otsu & Whang (2007) show that

\[
T_m^2 \overset{d}{\rightarrow} \chi^2_{\text{rank}(\Phi_m(\beta_0))}
\]

\[
C_g \overset{d}{\rightarrow} N (0, 1)
\]

\[
T_s^2 \overset{d}{\rightarrow} \chi^2_{\text{rank}(\Phi_s(\beta_0))}
\]

Therefore the CEL non nested test statistics follow standard distributions and as opposed to the CEL specification test of Tripathi & Kitamura (2003) discussed in the previous shows parametric convergence rates. To investigate the local power let

\[
H_{0n}^g = E \left[ g (z_i, \theta_n) | x_i \right] = \delta (x) / n^{1/2} \ a.s.
\]

denote a sequence of conditional local alternatives. Then it is possible to show that

\[
T_m^2 \overset{d}{\rightarrow} \chi^2_{\text{rank}(\Phi_m(\beta_0))} \left( \mu_m' \Phi_m (\beta_0)^{-g} \mu_m \right),
\]

\[
C_g \overset{d}{\rightarrow} N \left( \mu_c / \phi_c^{1/2}, 1 \right),
\]

\[
T_s^2 \overset{d}{\rightarrow} \chi^2_{\text{rank}(\Phi_s(\beta_0))} \left( \mu_s' \Phi_s (\beta_0)^{-g} \mu_s \right)
\]
where

\[
\begin{align*}
\mu_m &= -E \left[ I_i M(x_i, \theta_0, \beta_\star) J_i(\theta_0, \beta_\star) \Omega^g(\theta_0)^{-1} \delta(x_i) \right] + \\
& \quad H_m(\theta_0, \beta_\star) \Delta E[\delta_\psi(x_i)] \\
H_m(\theta_0, \beta_\star) &= E \left[ I_i M(x_i, \theta_0, \beta_\star) J_i(\theta_0, \beta_\star) \Omega^g(\theta_0)^{-1} G_i(\theta_0) \right] \\
\mu_c &= -2E \left[ I_i E \left[ h(\beta_\star) | x_i \right] \right]' \Omega^h(\beta)^{-1} J_i^h(\theta_0, \beta_\star) \Omega^g(\theta_0)^{-1} \delta(x_i) + \\
& \quad H_c(\theta_0, \beta_\star) \Delta E[\delta_\psi(x_i)] \\
H_c(\theta_0, \beta_\star) &= 2E \left[ I_i E \left[ h(\beta_\star) | x_i \right] \right]' \Omega^h(\beta)^{-1} J_i^h(\theta_0, \beta_\star) \Omega^g(\theta_0)^{-1} G_i(\theta_0) \\
\mu_s &= -E \left[ I_i G_i^h(\beta_\star) \Omega^h(\beta)^{-1} J_i^h(\theta_0, \beta_\star) \Omega^g(\theta_0)^{-1} \delta(x_i) \right] + \\
& \quad H_s(\theta_0, \beta_\star) \Delta E[\delta_\psi(x_i)] \\
H_s(\theta_0, \beta_\star) &= E \left[ I_i G_i^h(\beta_\star) \Omega^h(\beta)^{-1} J_i^h(\theta_0, \beta_\star) \Omega^g(\theta_0)^{-1} G_i(\theta_0) \right]
\end{align*}
\]

which shows that local power can be computed using standard noncentral distributions.


