Recent Developments in Empirical Likelihood and Related Methods

Paulo M.D.C. Parente[1,2] and Richard J. Smith[3,4]

[1]Department of Economics, University of Exeter, Exeter EX4 4ST, United Kingdom

[2]Centre for Research in Microeconomics, Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, United Kingdom

[3]Centre for Microdata Methods and Practice, University College London, and Institute of Fiscal Studies, London WC1E 7AE, United Kingdom

[4]Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, United Kingdom; email: rjs27@econ.cam.ac.uk

Annu. Rev. Econ. 2014. 6:77–102

First published online as a Review in Advance on February 5, 2014

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-080511-110925

JEL codes: C10, C30

Keywords

moment conditions, GMM, minimum discrepancy, generalized minimum contrast, generalized empirical likelihood

Abstract

This article reviews a number of recent contributions to estimation and inference for models defined by moment condition restrictions. The particular emphasis is on the generalized empirical likelihood class of estimators as an alternative to the generalized method of moments. Estimation methods for parameters defined through moment restrictions and their properties are described with tests of overidentifying moment restrictions and parametric hypotheses. Computational issues are discussed together with some proposals for their amelioration. Higher-order and other properties are also addressed in some detail. Models specified by conditional moment restriction models are considered, and the adaptation of these methods to weakly dependent data is discussed.

1. INTRODUCTION

Many, if not most, estimators commonly employed in empirical economic research may be motivated and formulated as the solution to a suitably defined set of moment restrictions. Indeed, the least squares estimator in the standard linear regression model is expressed in terms of the requirement that the sample covariance or correlation between regression residuals and the regressors is zero. Under correct specification, the parameters of the linear regression model are defined by the population counterpart of this condition. Hence, these moment estimators may be regarded as analog estimators based on the sample counterparts of population moment conditions (see Manski 1988). The maximum likelihood (ML) estimator is a further example that solves the likelihood equations (i.e., sets the first-order conditions or score vector to zero), with its population version being the zero expectation of the score (see, e.g., Goldberger 1991, chapter 12). A particular feature of these examples is that the number of parameters is identical to the number of moment restrictions; in other words, the model is just identified.

The models economists are often concerned with typically include explanatory variables that are themselves endogenous, for example, in the regression context when the dependent and one or more of the regressor variables or covariates are jointly determined. This may arise because of the omission of relevant variables, measurement error, or the economic model under investigation stipulating simultaneous determination. In such circumstances, in the regression context, the standard approach is to seek instruments or instrumental variables that are correlated with regressor variables but are uncorrelated with the regression error. The moment condition for estimation is then described by the covariance or correlation between the regression error and instruments being zero. The number of instruments, and thus moment conditions, may exceed the number of parameters to be estimated, yielding an overidentified model. Consequently, the sample analog of the moment restrictions can no longer be used directly for parameter estimation.

The standard approach to deal with an overidentified model minimizes a distance measure expressed in terms of the sample covariance between the regression errors and instruments. Of course, the resultant estimator will depend on the definition of distance adopted. For computational purposes, a weighted Euclidean distance is often convenient. A particular choice for the weight matrix results in the instrumental variable estimator introduced in S. Wright (1925), P.G. Wright (1928), and Reiersøl (1941, 1945) (see also Sargan 1958, 1959). Amemiya (1974) extends this approach to endogenous nonlinear regression models. Hansen (1982) considers a general setup with nonlinear moment restrictions and introduces the generalized method of moments (GMM) estimation procedure. GMM is asymptotically efficient if the weight matrix is chosen as the inverse of the variance matrix of the sample moments. Given an initial consistent parameter estimator, this variance matrix may be straightforwardly estimated, with the resultant two-step (2S)GMM estimator being asymptotically efficient.

There is now extensive simulation evidence that the statistical properties of 2SGMM for the sample sizes typically available to the empirical investigator can be quite unsatisfactory, which has stimulated intensive interest in other methods of estimation for moment condition models. Empirical likelihood (EL) was originally proposed by Owen (1988) for the estimation of a population mean and was introduced independently by Qin & Lawless (1994) and Imbens (1997) for models specified by the type of moment conditions considered by Hansen (1982). EL differs from GMM in that it is a nonparametric likelihood method of estimation based on a multinomial density formulation that incorporates the moment conditions

as restrictions. Consequently, EL displays obvious similarities to classical ML for fully parametric models.[1]

As is well known, classical ML possesses a number of optimality and other useful properties. First, ML is asymptotically efficient; in other words, the asymptotic variance of the ML estimator coincides with the Cramer-Rao lower bound for regular estimators. Second, bias-corrected ML is higher-order efficient (see Ghosh 1994). Third, the likelihood ratio (LR) test based on ML is optimal with respect to a large deviation optimality criterion (see Hoeffding 1965). Finally, Bartlett (1937) shows that the LR statistic for a test of a simple hypothesis may be scale corrected by what is now commonly referred to as a Bartlett correction, thereby ensuring a more rapid rate of convergence in distribution to the chi-squared distribution than the original LR statistic. This result has subsequently been extended to more general forms of parametric hypotheses (see, e.g., Cribari-Neto & Cordeiro 1996, and references therein).

Given its similarity to classical ML, perhaps it is unsurprising that EL also possesses some desirable and similar properties. Qin & Lawless (1994) and Imbens (1997) show that EL shares the asymptotically efficiency property of 2SGMM. Newey & Smith (2004) prove that bias-corrected EL is higher-order efficient. Kitamura (2001) and Kitamura et al. (2012) demonstrate that the EL criterion function test statistic for overidentifying moment conditions is optimal with respect to a large deviation optimality criterion. Chen & Cui (2006, 2007) and Matsushita & Otsu (2013) show that the EL criterion function test statistic for parametric restrictions and overidentifying moment conditions is Bartlett correctable.

Several alternatives to EL have also been proposed that share its first-order efficiency properties, including exponential tilting (ET) (see Kitamura & Stutzer 1997, Imbens et al. 1998) and the continuous updating estimator (CUE) (see Hansen et al. 1996). Corcoran (1998) introduces a class of estimators that minimizes a discrepancy measure between the empirical distribution function and the empirical distribution function constrained to satisfy the moment restrictions in the sample, the latter thereby constituting a saddle point problem. Several estimators belong to this class, including EL, ET, CUE, and those based on members of the Cressie-Read (1984) (CR) class of divergence measures proposed by Baggerly (1998). Kitamura (2007) considers a general class of *f*-divergence measures as defined by Csiszar (1963) and Ali & Silvey (1966) and defines the generalized minimum contrast (GMC) class of estimators.

Smith (1997, 2011) proposes a different class of estimators, motivated as a nonparametric adaptation to the moment condition setting of Chesher & Smith's (1997) approach, which develops LR tests for parametric moment conditions in the likelihood framework. Although it differs from GMC, the generalized empirical likelihood (GEL) class of estimators also requires the solution of a saddle point problem and includes EL, ET, CUE, and the CR class as special cases (see Smith 1997, Newey & Smith 2004). Newey & Smith (2004) prove that GEL contains a subclass of estimators that shares the same asymptotic bias as EL; if this GEL subclass is restricted further, bias-corrected GEL, similar to bias-corrected EL, is higher-order efficient.

[1]Evidence presented in the special section of the July 1996 issue of the *Journal of Business Economics and Statistics* indicates that 2SGMM may be severely biased and initiated interest in alternative estimation methods. Most papers concerned with moment condition models such as those considered in this review investigate finite sample mean and median estimator biases via Monte Carlo studies. Ramalho (2005) considers covariance structure models; although estimators cannot be ranked in terms of mean bias, median bias is lower for EL and related estimators than for 2SGMM. Guggenberger (2008) corroborates these results but observes that the standard deviation of EL and other estimators appears very large, suggesting that these estimators may not possess finite sample moments. Kitamura (2007) also studies mean and median bias of several estimators in a dynamic panel data model, a similar design to that analyzed by Imbens (2002); the conclusions are rather similar to those of Ramalho (2005) and Guggenberger (2008). Readers are also referred to Imbens & Spady (2005), Mittelhammer et al. (2005), and Newey et al. (2005).

The review is organized as follows. Section 2 describes the moment condition framework. Estimation methods for parameters defined through moment restrictions are given in Section 3. Tests of overidentifying moment restrictions and parametric hypotheses are presented in Section 4. In practice, although the large sample properties of (G)EL are attractive, computation may prove to be difficult in comparison to GMM because (G)EL solves a saddle point problem. Section 5 discusses this issue and some proposals for its amelioration. Higher-order and other properties of GEL are addressed in some detail in Section 6. Section 7 deals with conditional moment restriction models, whereas Section 8 considers how (G)EL, originally designed for cross-sectional context, may be suitably adapted for weakly dependent data. Section 9 concludes with a brief outline of some open research areas, together with a short discussion of other topics not addressed in this review because of space constraints. Useful additional references on (G)EL are Anatolyev & Gospodinov (2011), Imbens (2002), and Kitamura (2007).

2. MOMENT CONDITIONS

This section outlines the general framework used in this review. Several empirically relevant examples are provided as illustrations.

Let z denote a vector of d_z observable random variables. To describe the moment condition framework, let $g(z,\beta) = (g^1(z,\beta),\ldots,g^m(z,\beta))'$ denote the moment indicator vector, an m-vector of known functions of the data vector z and the p-vector of parameters β, which is of particular inferential interest to the investigator. The dimension m of the moment indicator vector is at least as great as p that of the parameter vector β; in other words, the model is either just identified, $m = p$, or overidentified, $m > p$. It is assumed that the moment conditions

$$\mathbb{E}\big[g(z,\beta)\big] = 0 \tag{1}$$

are uniquely satisfied when β takes the unknown true value β_0. Here $\mathbb{E}[\cdot]$ denotes expectation taken with respect to the distribution of z.

A number of well-known estimation problems fall within this setting.

> **Example 1 (maximum likelihood):** Suppose that z is distributed with probability density function $f(z, \beta)$ twice differentiable in β. It is assumed that although the function form of $f(z, \beta)$ is known, the parameter vector β_0 is not. The score vector $s(z, \beta)$ is the first-order derivative of the logarithm of the density $f(z, \beta)$; in other words, $s(z, \beta) = \partial \log f(z, \beta)/\partial\beta$, where $\log(\cdot)$ is the natural logarithm. It may be shown straightforwardly that
>
> $$\mathbb{E}\big[s(z,\beta_0)\big] = 0,$$
>
> where in this example $\mathbb{E}[\cdot]$ denotes expectation taken with respect to $f(z,\beta_0)$ (see, e.g., Goldberger 1991, p. 128).
>
> **Example 2 (quantile regression):** Let $z = (y, x')'$, where x is a random vector of dimension $d_x = d_z - 1$. The θ-quantile regression model is defined by the probability statement $\mathcal{P}\{y \le x'\beta_0\} = \theta$. The moment condition in Equation 1 defining the standard quantile regression estimator is given by
>
> $$\mathbb{E}\Big[x\big(\theta - I(y \le x'\beta_0)\big)\Big] = 0,$$
>
> where the indicator function $I(A) = 1$ if A is true and zero otherwise. Here $p = d_x$.

Examples 1 and 2 correspond to just-identified moment condition models, as $m = p$ in both cases. The next example allows for underidentification, $m < p$, just identification, $m = p$, and overidentification, $m > p$, as possibilities.

> **Example 3 (instrumental variables):** In this example, the observation vector z is redefined as $z = (y, x', w')'$, where w is a random d_w-vector of instruments or instrumental variables that satisfies the moment condition
>
> $$\mathbb{E}\left[w(y - x'\beta_0)\right] = 0,$$
>
> in other words, the standard instrument validity condition $\mathbb{E}[wu] = 0$ in which the regression error $u = y - x'\beta_0$ is uncorrelated with the vector of instruments w (see Equation 1). Here, as in Example 2, $p = d_x$, but now $m = d_w$.

3. ESTIMATION METHODS

In this section, z_i, $i = 1, \ldots, n$, denotes a random sample of data observations drawn from the distribution of z. Then, for a given β, the sample analog of the population expectation $\mathbb{E}[g(z, \beta)]$ is given by the sample mean $\hat{g}(\beta) = \sum_{i=1}^{n} g_i(\beta)/n$, where $g_i(\beta) = g(z_i, \beta)$, $i = 1, \ldots, n$. Additionally, let $\Omega(\beta) = \mathbb{E}[g(z, \beta)g(z, \beta)']$, and $\Omega = \Omega(\beta_0)$, the positive definite variance matrix of $g(z, \beta_0)$; the sample counterpart of $\Omega(\beta)$ is denoted by $\hat{\Omega}(\beta) = \sum_{i=1}^{n} g_i(\beta)g_i(\beta)'/n$. It is assumed in the following that the moment indicator vector $g(z, \beta)$ is first-order differentiable with respect to β with the consequent definitions of the full rank population Jacobian matrix $G = G(\beta_0)$, where $G(\beta) = \mathbb{E}[\partial g(z, \beta)/\beta']$, and the sample Jacobian $\hat{G}(\beta) = \sum_{i=1}^{n} G_i(\beta)/n$, where $G_i(\beta) = \partial g_i(\beta)/\partial\beta'$, $i = 1, \ldots, n$.

We use the common notation $\hat{\beta}$ for all efficient estimators of β_0 described below because, under suitable conditions, they share the same first-order large sample properties [i.e., consistency, $\hat{\beta} \xrightarrow{p} \beta_0$, root-$n$ asymptotic normality, and first-order asymptotic efficiency $n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma)$, where $\Sigma = (G'\Omega^{-1}G)^{-1}$ is the semiparametric efficiency lower bound] (Chamberlain 1987).

3.1. Efficient Generalized Method of Moments

That the moment condition $\mathbb{E}[g(z, \beta)] = 0$ (Equation 1) is satisfied uniquely at $\beta = \beta_0$ and that the sample mean $\hat{g}(\beta)$ should closely approximate the population mean $\mathbb{E}[g(z, \beta)]$ uniformly in β for all n sufficiently large suggest that an appropriate estimator of β_0 should minimize some measure of distance between $\hat{g}(\beta)$ and 0. These arguments motivate the GMM estimator originally proposed by Hansen (1982):

$$\tilde{\beta} = \arg\min_{\beta \in \mathcal{B}} \hat{g}(\beta)'\hat{W}^{-1}\hat{g}(\beta), \tag{2}$$

where $\mathcal{B}$ denotes the parameter space, and $\hat{W}$ is a positive semidefinite matrix such that $\hat{W}$ converges in probability to the positive definite matrix W.

Hansen (1982) shows that, under certain regularity conditions, the GMM estimator $\tilde{\beta}$ is consistent for β_0 and asymptotically normally distributed with asymptotic variance matrix given by $\operatorname{avar}[\tilde{\beta}] = \left(G'W^{-1}G\right)^{-1}G'W^{-1}\Omega W^{-1}G\left(G'W^{-1}G\right)^{-1}$; in other words, $n^{1/2}\left(\tilde{\beta} - \beta_0\right)$ converges in distribution to an $N(0, (G'W^{-1}G)^{-1}G'W^{-1}\Omega W^{-1}G(G'W^{-1}G)^{-1})$ distributed

random vector. Additionally, among the class of GMM estimators defined by Equation 2, the efficient GMM estimator sets $W = \Omega$. Given an initial consistent GMM estimate $\tilde{\beta}$ for β_0 (e.g., obtained by setting $\hat{W} = I_m$), then an efficient 2SGMM estimator results from replacing $\hat{W}$ with $\hat{\Omega}(\tilde{\beta})$ in Equation 2; that is,

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \hat{g}(\beta)' \hat{\Omega}(\tilde{\beta})^{-1} \hat{g}(\beta), \tag{3}$$

with asymptotic variance matrix $\Sigma = (G'\Omega^{-1}G)^{-1}$. The matrices G and Ω may be consistently estimated by $\hat{G}(\hat{\beta})$ and $\hat{\Omega}(\hat{\beta})$.

3.2. Empirical Likelihood

Owen (1988) originally proposed EL to define confidence regions for the population mean and differentiable functionals of the mean (see also Owen 1990, 2001). A generalization of EL to models specified by moment conditions of the form of Equation 1 is provided by Qin & Lawless (1994) and Imbens (1997).

Essentially, EL is a nonparametric generalization of parametric ML to the moment condition setting; indeed, if z is a vector of discrete distributed random variables, then EL is ML. EL treats the data as if they were discrete with probabilities π_i, $i = 1, \ldots, n$, assigned to each sample point, and similar to ML, EL estimates these probabilities so as to maximize the probability of observing the sample but subject to the imposition of the additional condition that the moment conditions are satisfied. To describe EL, we consider the multinomial log likelihood

$$\sum_{i=1}^{n} \log \pi_i. \tag{4}$$

The EL estimator of β_0 maximizes the criterion in Equation 4 subject to the unit simplex definitional constraint on the probabilities π_i, $i = 1, \ldots, n$, that is, nonnegativity $\pi_i \geq 0$, $i = 1, \ldots, n$, and unit summability $\sum_{i=1}^{n} \pi_i = 1$, together with the moment restriction $\sum_{i=1}^{n} \pi_i g_i(\beta) = 0$ (see Equation 1). After profiling out the probabilities π_i, $i = 1, \ldots, n$, and the Lagrange multiplier associated with the unit summability constraint, the EL criterion is

$$\mathcal{EL}_n(\beta, \lambda) = \sum_{i=1}^{n} \log\bigl(1 + \lambda' g_i(\beta)\bigr)/n, \tag{5}$$

where λ is the Lagrange multiplier associated with the sample moment constraint $\sum_{i=1}^{n} \pi_i g_i(\beta) = 0$. The EL estimator satisfies

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \sup_{\lambda \in \hat{\Lambda}_n(\beta)} \mathcal{EL}_n(\beta, \lambda), \tag{6}$$

where $\hat{\Lambda}_n(\beta) = \bigl\{\lambda : \lambda' g_i(\beta) > -1,\ i = 1, \ldots, n\bigr\}$ ensuring that the log function is well defined. The nonnegativity restriction is thus automatically satisfied because the estimated, typically referred to as implied or empirical, probabilities are given by

$$\hat{\pi}_i = \frac{1}{n\bigl(1 + \hat{\lambda}' g_i(\hat{\beta})\bigr)}, (i = 1, \ldots, n);$$

the Lagrange multiplier estimator is $\hat{\lambda} = \arg\max_{\lambda \in \hat{\Lambda}_n(\hat{\beta})} \sum_{i=1}^{n} \log\bigl(1 + \lambda' g_i(\hat{\beta})\bigr)$.

3.3. Minimum Discrepancy

Corcoran (1998) introduces a class of estimators based on the minimization of a discrepancy measure defined by

$$\mathcal{I}\left(\pi^n, \iota^n\right), \tag{7}$$

where $\pi^n = (\pi_1, \ldots, \pi_n)'$ and ι^n is an n-vector with all elements equal to the unrestricted empirical probabilities $1/n$. Minimum discrepancy (MD) estimators minimize Equation 7 with respect to π^n, and subject to $\pi_i \geq 0$, $i = 1, \ldots, n$, $\sum_{i=1}^n \pi_i = 1$ and $\sum_{i=1}^n \pi_i g_i(\beta) = 0$ (see Section 3.2). Several estimators belong to this class, in particular, EL, CUE (Hansen et al. 1996), ET (Kitamura & Stutzer 1997, Imbens et al. 1998), and the CR class of estimators (Cressie & Read 1984).

Kitamura (2007) suggests the use of f-divergence (see Csiszar 1963, Ali & Silvey 1966) to define the GMC class of estimators.[2] Here the discrepancy measure in Equation 7 is redefined as

$$\begin{aligned}\mathcal{I}(\pi^n, \iota^n) &= \sum_{i=1}^n \iota_i \phi(\pi_i/\iota_i) \\ &= \frac{1}{n}\sum_{i=1}^n \phi(n\pi_i),\end{aligned}$$

where $\iota_i = 1/n$, $i = 1, \ldots, n$, and $\phi(\cdot)$ is a convex function defined on the half line $[0, \infty)$ and continuous at zero such that $\phi(1) = 0$. For EL, one sees that $\phi(v) = -\log v$; for ET, $\phi(v) = v \log v$; for CUE, $\phi(v) = (v-1)^2/2$; and for CR, $\phi(v) = [v^{\tau+1} - v - \tau(v-1)]/[\tau(1+\tau)]$. The nonnegativity condition $\pi_i \geq 0$, $i = 1, \ldots, n$, essential for their interpretation as empirical probabilities, is often ignored in practice for CUE and CR, although it must be imposed for EL (see Section 3.2), and is automatically satisfied for ET. Kitamura (2007) shows that the dual of Equation 7 is given by

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \sup_{\gamma, \lambda} \sum_{i=1}^n \left(\gamma - \frac{1}{n}\phi^*\left(\gamma + \lambda' g_i(\beta)\right)\right),$$

where $\phi^*(v) = \sup_x [xv - \phi(x)]$ is the Legendre transform of $\phi(\cdot)$. It follows that for EL, $\phi^*(v) = -1 - \log(-v)$; for ET, $\phi^*(v) = \exp(v) - 1$; for CUE, $\phi^*(v) = v^2/2 + v$; and for CR, $\phi^*(v) = (\tau v + 1)^{1/\tau+1}/(\tau+1) - 1/(\tau+1)$.

3.4. Generalized Empirical Likelihood

GEL (introduced in Smith 1997; see also Smith 2011) differs in general from MD and GMC.[3] Although it is not explicitly defined in terms of a program based on empirical probabilities, GEL also includes EL, ET, CUE, and the CR class of estimators as special cases.

The GEL class of estimators is defined as follows. Let

[2]The f-divergence between two discrete probability distributions $p = \{p_1, p_2, \ldots\}$ and $q = \{q_1, q_2, \ldots\}$ is defined by $D_\phi(p, q) = \sum_{i=1}^\infty p_i \phi(q_i/p_i)$, where the function $\phi(\cdot)$ is defined below.

[3]Smith (1997, 2011) motivates GEL as a nonparametric generalization to the moment condition context of the approach taken by Chesher & Smith (1997) in a fully parametric likelihood setting. Chesher & Smith (1997) propose LR test statistics for implied moment conditions in which the likelihood augments the null hypothesis parametric density multiplicatively by a function of a weighted version of the moment indicators underpinning the implied moment conditions.

$$\hat{P}_n^{\rho}(\beta,\lambda) = \sum_{i=1}^{n}\left[\rho\left(\lambda' g_i(\beta)\right) - \rho_0\right]\Big/ n,$$

where the function $\rho(\cdot)$ is concave on its domain $\mathcal{V}$, an open interval containing zero, with derivatives $\rho_j(v) = \partial^j \rho(v)/\partial v^j$, $\rho_j(0) = \rho_j$, $j = 0, 1, \ldots$, normalized without loss of generality as $\rho_1 = \rho_2 = -1$. The GEL estimator of β_0 is given by

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} \sup_{\lambda\in\hat{\Lambda}_n(\beta)} \hat{P}_n^{\rho}(\beta,\lambda), \tag{8}$$

where $\hat{\Lambda}_n(\beta) = \left\{\lambda : \lambda' g_i(\beta) \in \mathcal{V}, i = 1, \ldots, n\right\}$ with the Lagrange multiplier-like estimator $\hat{\lambda} = \arg\max_{\lambda\in\hat{\Lambda}_n(\hat{\beta})} \hat{P}_n^{\rho}\left(\hat{\beta},\lambda\right)$ the first-order condition for which imposes the sample moment constraint $\sum_{i=1}^{n}\hat{\pi}_i g_i\left(\hat{\beta}\right) = 0$ (see Section 3.2). The implied GEL empirical probabilities $\hat{\pi}_i$, $i = 1, \ldots, n$, are

$$\hat{\pi}_i = \frac{\rho_1\left(\hat{\lambda}'\hat{g}_i\right)}{\sum_{j=1}^{n}\rho_1\left(\hat{\lambda}'\hat{g}_j\right)}, (i = 1, \ldots, n), \tag{9}$$

summing to one by construction, but are typically not all nonnegative in finite samples, where $\hat{g}_i = g_i\left(\hat{\beta}\right)$, $i = 1, \ldots, n$.[4] For any function $a(z,\beta)$, a semiparametrically efficient estimator of the moment $\mathbb{E}[a(z,\beta_0)]$ is formed from the empirical probabilities as $\sum_{i=1}^{n}\hat{\pi}_i a\left(z_i,\hat{\beta}\right)$ (see Brown & Newey 1998).

As noted above, GEL does not coincide with MD or GMC. Because GEL is the dual of the *f*-divergence program for the CR class that includes EL, ET, and CUE (Smith 1997, Newey & Smith 2004), GEL therefore yields the same estimators for this class (see Kitamura 2007). Newey & Smith (2004), Kitamura (2007), and Smith (2007c) note that this result holds if the inverse of $\phi(\cdot)$ defining the GMC class in Equation 8 is homogeneous. GEL includes EL [$\rho(v) = \log(1+v)$ and $\mathcal{V} = (-1, +\infty)$], ET [$\rho(v) = -\exp(v)$], CUE [$\rho(v) = -(v+1)^2/2$], and CR [$\rho(v) = -(1+\tau v)^{(1+\tau)/\tau}/(1+\tau)$].

Newey & Smith (2004) obtain the joint limit distribution of $\hat{\beta}$ and $\hat{\lambda}$ as

$$n^{1/2}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\lambda}\end{pmatrix} \xrightarrow{d} N\left(0, \operatorname{diag}(\Sigma, P)\right),$$

where $P = \Omega^{-1} - \Omega^{-1}G\Sigma G'\Omega^{-1}$. Indeed, a first-order asymptotically equivalent Lagrange multiplier-like estimator is obtained from the program $\hat{\lambda} = \arg\max_{\lambda\in\hat{\Lambda}_n\left(\hat{\beta}\right)} \hat{P}_n^{\rho}\left(\hat{\beta},\lambda\right)$ for any estimator $\hat{\beta}$ first-order asymptotically equivalent to GEL, for example, efficient 2SGMM, MD, or GMC, with associated empirical probabilities consequently defined as in Equation 9 (see Brown & Newey 2002).

4. TESTS

4.1. Overidentifying Moment Conditions

An important hypothesis of interest to empirical researchers is whether the moment conditions in Equation 1 hold. Consider the null hypothesis

[4]The shrinkage estimators $\tilde{\pi}_i = \left(\hat{\pi}_i + n^{-1}\varepsilon_n\right)/(1+\varepsilon_n)$, $i = 1, \ldots, n$, where $\varepsilon_n = -n\min[\min_{1\le i\le n}\hat{\pi}_i, 0]$, deal with this problem without affecting the large sample analysis (see Antoine et al. 2007, equations 2.8 and 2.9, p. 466). Empirical probabilities are given for EL by Owen (1988), for ET by Kitamura & Stutzer (1997), for quadratic $\rho(\cdot)$ by Back & Brown (1993), and for the general case by Brown & Newey (2002).

$$H_0 : \mathbb{E}\big[g(z,\beta)\big] = 0 \text{ for some } \beta \in \mathcal{B}$$

and the associated alternative hypothesis

$$H_1 : \mathbb{E}\big[g(z,\beta)\big] \neq 0 \text{ for all } \beta \in \mathcal{B}.$$

Hansen (1982) considers an overidentified setting in which the number of moment restrictions m exceeds the number of parameters p and proposes what is now commonly known as the $\mathcal{J}$-statistic to test H_0 against H_1, that is, the optimized efficient GMM criterion (see Equation 3)

$$\mathcal{J}_n = n\hat{g}\big(\hat{\beta}\big)'\hat{\Omega}\big(\tilde{\beta}\big)^{-1}\hat{g}\big(\hat{\beta}\big).$$

Hansen (1982) demonstrates that if the moment restrictions in Equation 1 hold (and thus H_0 is true), then $\mathcal{J}_n$ has a limiting chi-squared distribution with $m - p$ degrees of freedom with consequent asymptotic α-level critical or rejection region $\big\{\mathcal{J}_n > c^{\alpha}_{m-p}\big\}$, where $\mathcal{P}\big\{\chi^2_{m-p} > c^{\alpha}_{m-p}\big\} = \alpha$.

Although the $\mathcal{J}$-statistic is straightforward to compute, several simulation studies have cast doubt on whether its asymptotic properties are a useful guide to its performance in finite samples. Alternative test statistics based on GEL and associated criteria have also been proposed (see Kitamura & Stutzer 1997; Smith 1997, 2000, 2011; Imbens et al. 1998; Newey & Smith 2004). These statistics parallel the classical trinity of LR, Lagrange multiplier, and score statistics, namely, an LR form of statistic

$$\mathcal{LR}_n = 2n\hat{P}^{\rho}_n\big(\hat{\beta},\hat{\lambda}\big), \tag{10}$$

a Lagrange multiplier statistic

$$\mathcal{LM}_n = n\hat{\lambda}'\hat{\Omega}\big(\hat{\beta}\big)\hat{\lambda},$$

and a score statistic

$$\mathcal{S}_n = n\hat{g}\big(\hat{\beta}\big)'\hat{\Omega}\big(\hat{\beta}\big)^{-1}\hat{g}\big(\hat{\beta}\big),$$

where $\hat{\beta}$ and $\hat{\lambda}$ denote GEL or first-order equivalent estimators (see Section 3). All three forms of test statistic are asymptotically equivalent to the $\mathcal{J}$-statistic with a chi-squared limiting distribution with $m - p$ degrees of freedom if the moment restrictions in Equation 1 (and H_0) hold.

4.2. Parametric Restrictions

Owen (1990) suggests an EL-based LR-type statistic to test the simple null hypothesis $H_0 : \beta_0 = \beta^0$ against the alternative $H_0 : \beta_0 \neq \beta^0$, where β^0 is a known p-vector of constants, when $g(z, \beta) = z - \beta$ and β_0 denotes the population mean (i.e., $m = p$):

$$\mathcal{LR}^r_n = 2n\hat{P}^{\rho}_n\Big(\beta^0,\ \hat{\lambda}\big(\beta^0\big)\Big), \tag{11}$$

where $\hat{P}^{\rho}_n(\beta,\lambda)$ is the EL criterion $\mathcal{EL}_n(\beta,\lambda)$ (Equation 5), $\hat{\lambda}(\beta) = \arg\max_{\lambda\in\hat{\Lambda}_n(\beta)}\hat{P}^{\rho}_n(\beta,\lambda)$, and $\hat{\Lambda}_n(\beta) = \{\lambda : \lambda' g_i(\beta) > -1, i = 1,\ldots,n\}$. If $H_0 : \beta_0 = \beta^0$ is true, $\mathcal{LR}^r_n$ converges in distribution

to a chi-squared random variable with p degrees of freedom (see Owen 1990, theorem 1, p. 91).[5]

This framework may be straightforwardly generalized to enable the construction of tests of functions of β_0; that is,

$$H_0 : r(\beta_0) = 0 \text{ against } H_1 : r(\beta_0) \neq 0,$$

where $r(\cdot)$ is a known differentiable s-vector of functions with dimension $s \leq p$; the moment restrictions in Equation 1 $\mathbb{E}[g(z, \beta_0)] = 0$ are maintained throughout.

Several classical-like GEL statistics have been proposed for testing $H_0 : r(\beta_0) = 0$ against $H_1 : r(\beta_0) \neq 0$, which are GEL counterparts of those suggested for GMM in Newey & West (1987). Let $R(\beta) = \partial r(\beta)/\partial \beta'$ with $R = R(\beta_0)$ of full rank s and define the restricted parameter space $\mathcal{B}^r = \{\beta \in \mathcal{B} : r(\beta) = 0\}$. The restricted GEL estimator is given by

$$\hat{\beta}^r = \arg\min_{\beta \in \mathcal{B}^r} \sup_{\lambda \in \hat{\Lambda}_n(\beta)} \hat{P}_n^{\rho}(\beta, \lambda),$$

and $\hat{\lambda}^r = \arg\max_{\lambda \in \hat{\Lambda}_n\left(\hat{\beta}^r\right)} \hat{P}_n^{\rho}\left(\hat{\beta}^r, \lambda\right)$. Let $\hat{R}$, $\hat{G}$, and $\hat{\Omega}$ be H_0-consistent estimates of G and Ω, respectively [e.g., $\hat{R} = R\left(\hat{\beta}\right)$, $\hat{G} = \hat{G}\left(\hat{\beta}\right)$, and $\hat{\Omega} = \hat{\Omega}\left(\hat{\beta}\right)$]. Define $\hat{\Sigma} = \left(\hat{G}'\hat{\Omega}^{-1}\hat{G}\right)^{-1}$ and $\hat{H} = \hat{\Sigma}\hat{G}'\hat{\Omega}^{-1}$.

GEL statistics for testing $H_0 : r(\beta_0) = 0$ against $H_1 : r(\beta_0) \neq 0$ are (see Smith 1997, 2000, 2011) an LR statistic,

$$\mathcal{LR}_n^r = 2n\left(\hat{P}_n^{\rho}\left(\hat{\beta}^r, \hat{\lambda}^r\right) - \hat{P}_n^{\rho}\left(\hat{\beta}, \hat{\lambda}\right)\right); \tag{12}$$

a Wald statistic,

$$\mathcal{W}_n^r = nr\left(\hat{\beta}\right)'\left(\hat{R}\hat{\Sigma}\hat{R}'\right)^{-1} r\left(\hat{\beta}\right);$$

a Lagrange multiplier statistic,

$$\mathcal{LM}_n^r = n\left(\hat{\lambda}^r - \hat{\lambda}\right)'\hat{\Omega}\left(\hat{\lambda}^r - \hat{\lambda}\right);$$

and a score statistic,

$$\mathcal{S}_n^r = n\hat{g}\left(\hat{\beta}^r\right)'\hat{H}'\hat{R}'\left(\hat{R}\hat{\Sigma}\hat{R}'\right)^{-1}\hat{R}\hat{H}\hat{g}\left(\hat{\beta}^r\right).$$

Under standard conditions, if $H_0 : r(\beta_0) = 0$ holds, all the above statistics are asymptotically equivalent and have a limiting chi-squared distribution with s degrees of freedom with [e.g., for $\mathcal{LR}_n^r$ (Equation 12)] asymptotic α-level critical or rejection region $\{\mathcal{LR}_n^r > c_s^{\alpha}\}$.

[5] Hjört et al. (2009) allow the number of parameters p to diverge and approach infinity with the sample size n. If the moment indicator vector $g(z, \beta)$ is uniformly bounded, then the critical region $\left\{2n\hat{P}_n^{\rho}(\beta_0, \hat{\lambda}(\beta_0)) > c_p^{\alpha}\right\}$ is still valid for an asymptotic α-level test provided that $p^3/n \to 0$ (see Hjört et al. 2009, theorem 4.1, p. 1098).

4.3. Nuisance Parameters

In some cases, the moment restrictions in Equation 1 may depend on nuisance parameters, such as unknown functions or parameters that are not defined explicitly by the moment restrictions but can be estimated using extraneous information. To describe this setting, we redefine the moment restrictions in Equation 1 as

$$\mathbb{E}\big[g(z,\beta_0,h_0)\big]=0,$$

where, as before, the vector of moment indicators is known up to β_0 but now includes the unknown vector of functions or parameters h_0.

The following examples are taken from Hjört et al. (2009), who adopt a plug-in approach with an estimator $\hat{h}$, for example, substituted for the unknown h_0 in the moment indicator vector.

Example 4 (symmetry): The interest is in testing for the symmetry of the distribution of a random variable z around its median h_0. Consider a particular point z_0 in the range of z. Then symmetry at z_0 is expressed by the restriction

$$F_z(z_0)=1-F_z(2h_0-z_0),$$

where $F_z(\cdot)$ is the distribution function of z. Define $\beta_0=F_z(z_0)$. Then the moment restrictions

$$\mathbb{E}\big[I(z\le z_0)-\beta_0\big]=0,$$

$$\mathbb{E}\big[I(z>2h_0-z_0)-\beta_0\big]=0$$

are equivalent to symmetry at z_0 as defined above. The unknown population parameter h_0 is estimated by the sample median of z.

Example 5 (nonparametric regression error distribution): Let $z=(y,x)'$, where y and x are scalar random variables. Consider the nonparametric regression model

$$y=h_0(x)+u,$$

where $h_0(\cdot)$ is the unknown conditional mean function $\mathbb{E}[y|x]$ of y given x; the covariate x and the regression error u are assumed to be independent. Let $F_u(\cdot)$ denote the unknown distribution function of the regression error u. Given a fixed z_0, the distribution function of u at z_0 [i.e., $\beta_0=F_u(z_0)$] is of interest. The associated moment condition is then given by

$$\mathbb{E}\Big[I\big(y-h_0(x)\le z_0\big)-\beta_0\Big]=0.$$

A standard estimator for the unknown conditional mean function $h_0(\cdot)$ is the Nadaraya-Watson nonparametric estimator

$$\hat{h}(x)=\sum_{i=1}^{n}w_iy_i,$$

where $w_i=\mathcal{K}_i/\sum_{j=1}^{n}\mathcal{K}_j$, with $\mathcal{K}_i=\mathcal{K}\big((x-x_i)/b_n\big)$, $i=1,\ldots,n$, $k(\cdot)$ a symmetric positive kernel function, and b_n a bandwidth parameter.

Hjört et al. (2009) study the EL-based criterion function statistic $2n\hat{P}_n^{\rho}\big(\beta^0,\hat{\lambda}(\beta^0)\big)$ (Equation 11) for the hypothesis $H_0:\beta_0=\beta^0$ based on the moment indicator vector $g\big(z,\beta,\hat{h}\big)$. The limiting

distribution of the EL-based statistic is nonstandard for the above and other problems but can be approximated using bootstrap methods.

5. COMPUTATION

Because the GEL objective function may be highly nonlinear after profiling out the auxiliary parameter vector λ, there may be severe difficulties associated with the computation of the GEL estimator of β_0. Imbens & Spady (2002), Mittelhammer et al. (2005), and Kitamura (2007) advocate the following computational method. The profile GEL criterion function is defined by

$$\begin{aligned}\hat{P}_n^{\rho}(\beta) &= \hat{P}_n^{\rho}\left(\beta, \hat{\lambda}(\beta)\right) \\ &= \max_{\lambda \in \hat{\Lambda}_n(\beta)} \hat{P}_n^{\rho}(\beta, \lambda).\end{aligned}$$

Minimization of $\hat{P}_n^{\rho}(\beta)$ over $\beta \in \mathcal{B}$ constitutes the outer-loop problem, which may be complex because of nonlinearity. The Davidon-Fletcher-Powell (Imbens & Spady 2002) or the Nelder-Mead simplex (Mittelhammer et al. 2005) methods may be efficacious for the minimization of $\hat{P}_n^{\rho}(\beta)$. The latter method is possibly more preferable because neither the computation of the gradient nor the Hessian of $\hat{P}_n^{\rho}(\beta)$ nor numerical approximations to them are required, which may sometimes be problematic in practice given the dependence of $\hat{P}_n^{\rho}(\beta)$ on the inner-loop problem.

The inner-loop problem concerns the determination of $\hat{\lambda}(\beta)$ for a given $\beta \in \mathcal{B}$; in other words,

$$\hat{\lambda}(\beta) = \arg\max_{\lambda \in \hat{\Lambda}_n(\beta)} \hat{P}_n^{\rho}(\beta, \lambda). \tag{13}$$

Computation of $\hat{\lambda}(\beta)$ is relatively simple, as $\hat{P}_n^{\rho}(\beta, \lambda)$ is strictly concave on $\mathcal{V}$ and can easily be achieved by Newton or related methods because the first-order derivative and Hessian of $\hat{P}_n^{\rho}(\beta, \lambda)$ are straightforwardly obtained as

$$\frac{\partial \hat{P}_n^{\rho}(\beta, \lambda)}{\partial \lambda} = \sum\nolimits_{i=1}^{n} \rho_1\left(\lambda' g_i(\beta)\right) g_i(\beta)/n,$$

and

$$\frac{\partial^2 \hat{P}_n^{\rho}(\beta, \lambda)}{\partial \lambda \partial \lambda'} = \sum\nolimits_{i=1}^{n} \rho_2\left(\lambda' g_i(\beta)\right) g_i(\beta) g_i(\beta)'/n.$$

Because $v > -1$ is required for EL, where $\rho(v) = \log(1 + v)$, Kitamura (2007) suggests solving Equation 13 subject to the restriction $\lambda' g_i(\beta) \geq -1 + \delta$ for some small $\delta > 0$, $i = 1, \ldots, n$. An alternative (see Owen 2001, equation 12.3, p. 235) replaces the logarithmic function by

$$\begin{cases} \log(x) & \text{if } x \geq \xi \\ \log(\xi) - 1.5 + 2(x/\xi) - 0.5(x/\xi)^2 & \text{if } x < \xi \end{cases}, \tag{14}$$

which has support given by the real line for any small number $\xi > 0$. For ET, where $\rho(v) = -\exp(v)$, Imbens et al. (1998) use a penalty function approach that consists of solving the following problem:

$$\max_{\beta\in\mathcal{B},\lambda} K(\beta,\lambda)-\frac{1}{2}AK_\lambda(\beta,\lambda)'W^{-1}K_\lambda(\beta,\lambda),$$

where $K(\beta,\lambda)=\log\left(\sum_{i=1}^n \exp\left(\lambda' g_i(\beta)\right)/n\right)$, $K_\lambda(\beta,\lambda)=\partial K(\beta,\lambda)/\partial\lambda$, W is a positive definite matrix, and A is a positive scalar that can take large values. Imbens et al. (1998) choose W as $W=K_{\lambda\lambda}\left(\tilde{\beta},\tilde{\lambda}\right)+K_\lambda\left(\tilde{\beta},\tilde{\lambda}\right)K_\lambda\left(\tilde{\beta},\tilde{\lambda}\right)'$, where $K_{\lambda\lambda}(\beta,\lambda)=\partial^2 K(\beta,\lambda)/\partial\lambda\partial\lambda'$, and $\left(\tilde{\beta},\tilde{\lambda}\right)$ are initial estimates [e.g., the initial root-n consistent estimator $\tilde{\beta}$ used in 2SGMM and $\tilde{\lambda}=\hat{\lambda}\left(\tilde{\beta}\right)$ obtained from Equation 13].[6]

An additional computational issue is that GEL requires the associated empirical probabilities to be chosen so that, not only $\pi_i\geq 0$, $i=1,\ldots,n$, and $\sum_{i=1}^n\pi_i=1$ hold, but also the first-order condition $\sum_{i=1}^n\pi_i g_i(\beta)=0$ is automatically satisfied (i.e., $0\in\left\{\sum_{i=1}^n\pi_i g_i(\beta)|\pi_i\geq 0, i=1,\ldots,n, \sum_{i=1}^n\pi_i=1\right\}$ the convex hull of $\left\{g_i(\beta)\right\}_{i=1}^n$). In finite samples, this may not be possible for particular data configurations, $\left\{g_i(\beta)\right\}_{i=1}^n$. An attractive solution, adjusted EL, proposed recently by Chen et al. (2008) and Liu & Chen (2010) is to add a new data point to $\left\{g_i(\beta)\right\}_{i=1}^n$ defined by $g_{n+1}(\beta)=-a_n\hat{g}(\beta)$, where $\{a_n\}$ denotes a positive sequence, that thereby guarantees that zero is in the convex hull of $g_i(\beta)$, $i=1,\ldots,n+1$, because $\hat{g}(\beta)$ and $g_{n+1}(\beta)$ lie in this set and have opposite sign. In addition, the nonnegativity of the empirical probabilities $\hat{\pi}_i$, $i=1,\ldots,n$, in Equation 9 may not hold without explicitly imposing this condition, although for large samples $\hat{\pi}_i\geq 0, i=1,\ldots,n$, with probability close to one if the moment restrictions in Equation 1 are valid (footnote 4 suggests another approach).

6. HIGHER-ORDER PROPERTIES

6.1. Asymptotic Bias

Newey & Smith (2004) investigate the asymptotic bias of efficient 2SGMM and GEL using the stochastic expansion

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)=\psi_n+Q_{1,n}/\sqrt{n}+Q_{2,n}/n+Q_{3,n}/n^{3/2},$$

where the quantities $Q_{j,n}$, $j=1,\ldots,3$, are random vectors that are bounded in probability, and ψ_n has zero mean converging in distribution to an $N(0,\Sigma)$ distributed random vector.

The asymptotic bias of 2SGMM and GEL to order $O(n^{-1})$ requires only the analysis of the behavior of the order $1/\sqrt{n}$ term (i.e., $Q_{1,n}$), because $\mathbb{E}[\psi_n]=0$, and to this order $Q_{2,n}$ also has mean zero. More precisely, the asymptotic bias of 2SGMM and GEL is defined as $\text{abias}\left[\hat{\beta}\right]=\mathbb{E}[Q_{1,n}]/n$. In general, the $O(n^{-1})$ bias of 2SGMM and GEL may be decomposed into four terms; for efficient 2SGMM,

$$\text{abias}\left[\hat{\beta}\right]=B_I+B_G+B_\Omega+B_W; \tag{15}$$

for GEL,

$$\text{abias}\left[\hat{\beta}\right]=B_I+(1+\rho_3/2)B_\Omega. \tag{16}$$

Each term in Equations 15 and 16 has an interpretation. The first term, B_I, is the asymptotic bias of an efficient GMM estimator based on the infeasible optimal combination of moment condition

[6]Chaussé (2010) discusses computation of GMM and GEL using R. Stata code for EL is provided by Y. Kitamura at **http://kitamura.sites.yale.edu/**.

indicators $G'\Omega^{-1}g(z,\beta)$ with first-order conditions $G'\Omega^{-1}\hat{g}(\beta)=0$. The term B_G arises from the (implicit) estimation of the population Jacobian matrix G, whereas the estimation of the moment variance matrix Ω produces B_Ω. The term B_W appears because of the use of the preliminary consistent estimator $\tilde{\beta}$ for β_0 in efficient 2SGMM and is thus absent for GEL.[7]

Newey & Smith (2004) show that not only is the term B_W absent for GEL, but the Jacobian contribution B_G also vanishes. Clearly, if the third-order derivative $\rho_3=-2$, the moment variance term B_Ω disappears from Equation 16. Indeed, EL satisfies this condition.[8] To illustrate these results, Newey & Smith (2004, section 4.1, pp. 229–30) consider a model defined through the conditional moment restriction $\mathbb{E}[u(z,\beta_0)|x]=0$, where $u(z,\beta_0)$ is a scalar function; estimation of β_0 uses the unconditional moment indicators $g(z,\beta_0)=q(x)\times u(z,\beta_0)$ with the unconditional moment restrictions $\mathbb{E}[g(z,\beta_0)]=0$ of Equation 1, where $q(\cdot)$ is an m-vector of functions. Interestingly, in contradistinction to 2SGMM, EL asymptotic bias does not increase with the number of moment conditions m.

Schennach (2007) reconsiders exponentially tilted EL [EL(ET)], which incorporates the ET empirical probabilities into the EL objective function and was originally proposed by Jing & Wood (1996) and Corcoran (1998) for the population mean case. EL(ET) has the same asymptotic bias as EL, is also higher-order efficient, and possesses desirable properties when the moment conditions in Equation 1 are misspecified. An alternative approach is to embed the GEL rather than ET empirical probabilities, namely, $\hat{\pi}_i(\beta)=\rho_1\left(\hat{\lambda}(\beta)'g_i(\beta)\right)\big/\sum_{j=1}^{n}\rho_1\left(\hat{\lambda}(\beta)'g_j(\beta)\right)$, $i=1,\ldots,n$, where $\hat{\lambda}(\beta)=\arg\max_{\lambda\in\hat{\Lambda}_n(\beta)}\hat{P}_n^{\rho}(\beta,\lambda)$, into the EL objective function, yielding the EL (GEL) estimator (see Smith 2007c)

$$\hat{\beta}=\arg\max_{\beta\in\mathcal{B}}\sum_{i=1}^{n}\log\hat{\pi}_i(\beta)/n.$$

Note that the empirical probabilities $\hat{\pi}_i(\beta)$, $i=1,\ldots,n$, must be positive, a property satisfied by, for example, members of the CR family with $\tau\leq 0$. EL(GEL) has the same asymptotic bias as EL(ET) and thus EL, but whether bias-corrected EL(GEL) is also higher-order efficient remains to be proven.

To alleviate the potential computational difficulties associated with the GEL class of estimators (see Section 5), Fan et al. (2011) introduce an iterative scheme that yields estimators with the same asymptotic bias as GEL, which may be regarded as a development for the moment condition context of the classical likelihood approach in Robinson (1988b). Define the scalar function $k(v)=(\rho_1(v)+1)/v$, where $v\neq 0$, and $k(0)=-1$ (see Newey & Smith 2004, theorem 2.3, p. 224). Let $\hat{\beta}^{0}$ denote any root-n consistent initial estimator of β_0, for example, GMM with $\hat{W}=I_m$, with, for $j>0$, $\hat{\beta}^{(j)}$ the j-th iterate. Also let $\hat{\lambda}^{(j)}=\arg\max_{\lambda\in\hat{\Lambda}_n\left(\hat{\beta}^{(j)}\right)}\hat{P}_n^{\rho}\left(\hat{\beta}^{(j)},\lambda\right)$. Define $\hat{\pi}_i^{(j)}=\hat{\pi}_i\left(\hat{\beta}^{(j)}\right)$, $\hat{g}_i^{(j)}=g_i\left(\hat{\beta}^{(j)}\right)$, $i=1,\ldots,n$, $\hat{g}^{(j)}=\hat{g}\left(\hat{\beta}^{(j)}\right)$, $\hat{G}^{(j)}=\hat{G}\left(\hat{\beta}^{(j)}\right)$, $\tilde{G}^{(j)}=$

[7]Let a be the $m\times 1$ vector such that $a_j=tr\left(\Sigma\mathbb{E}\left[\partial^2 g_{ij}(\beta_0)/\partial\beta\partial\beta'\right]\right)/2$, $j=1,\ldots,m$, where $g_{ij}(\beta)$ denotes the j-th element of $g_i(\beta)$, $g_i=g_i(\beta_0)$, and $G_i=G_i(\beta_0)$. Also let $H_W=(G'W^{-1}G)^{-1}G'W^{-1}$, $H=\Sigma G'\Omega^{-1}$, $\overline{\Omega}_{\beta_j}=\mathbb{E}\left[\partial\left(g_i(\beta_0)g_i(\beta_0)'\right)/\partial\beta_j\right]$, and e_j the j-th unit vector. Then $B_I=H(-a+\mathbb{E}[G_iHg_i])/n$, $B_G=-\Sigma\mathbb{E}[G_i'Pg_i]/n$, $B_\Omega=H\mathbb{E}[g_ig_i'Pg_i]/n$, and $B_W=-H\sum_{j=1}^{p}\overline{\Omega}_{\beta_j}(H_W-H)'e_j/n$.

[8]Bias-corrected EL is higher-order efficient in the sense that it has the least higher-order asymptotic variance (see Newey & Smith 2004, theorem 6.1, p. 234). Moreover, any bias-corrected GEL estimator with the same derivatives ρ_j up to order four as EL, in particular, $\rho_3=-2$ and $\rho_4=-6$, shares this property.

$\sum_{i=1}^{n} \hat{\pi}_i^{(j)} G_i\left(\hat{\beta}^{(j)}\right)$, $\hat{\Omega}^{(j)} = \hat{\Omega}\left(\hat{\beta}^{(j)}\right)$, and $\tilde{\Omega}^{(j)} = \sum_{i=1}^{n} \hat{k}_i^{(j)} g_i\left(\hat{\beta}^{(j)}\right) g_i\left(\hat{\beta}^{(j)}\right)'$, where $\hat{k}_i^{(j)} = k\left(\hat{\lambda}^{(j)\prime} \hat{g}_i^{(j)}\right) \Big/ \sum_{i=1}^{n} k\left(\hat{\lambda}^{(j)\prime} \hat{g}_i^{(j)}\right)$, $i = 1, \ldots, n$. Then the j-th iterate $\hat{\beta}^{(j)}$ is defined as the solution to

$$\hat{G}^{(j)\prime}\left[\hat{\Omega}^{(j-1)}\right]^{-1} \hat{g}^{(j)} = \left(\hat{G}^{(j-1)\prime}\left[\hat{\Omega}^{(j-1)}\right]^{-1} - \tilde{G}^{(j-1)\prime}\left[\tilde{\Omega}^{(j-1)}\right]^{-1}\right) \hat{g}^{(j-1)},$$

which may be interpreted as the recentered 2SGMM first-order conditions with weight matrix $\hat{\Omega}^{(j-1)}$. Fan et al. (2011, theorems 4.1 and 4.2, p. 272) show that, for $j > 0$, $\hat{\beta}^{(j)}$ is asymptotically equivalent to the corresponding GEL estimator and thereby asymptotically efficient. If $\hat{\beta}^{(0)}$ is itself asymptotically efficient, then $\hat{\beta}^{(j)}$ has the same asymptotic bias as GEL for $j > 0$.

6.2. Bartlett Correction

The asymptotic distributions of the test statistics described in Section 4 are approximately chi-squared for sufficiently large sample sizes. As is widely recognized, the distribution of a statistic for the sample sizes typically available in practice may differ substantially from that predicted by large sample theory.

For fully parametric problems addressed by classical likelihood theory, a simple scale transformation of the (log) LR statistic, known as a Bartlett correction, results in an improved accuracy of the asymptotic chi-squared distribution theory for the finite sample behavior of the transformed statistic (see, e.g., Bartlett 1937, Cribari-Neto & Cordeiro 1996).[9] Similar results for the moment condition context are scarce (see, e.g., Owen 2001, pp. 249–51).

DiCiccio et al. (1991) discuss the smooth function model in which the moment indicator vector takes the form $g(z, \beta) = z - \beta$, where β_0 is the population mean, and consider the null hypothesis $H_0 : \theta(\beta_0) = \theta^0$ against the alternative $H_1 : \theta(\beta_0) \neq \theta^0$, where θ^0 is a known s-vector of constants and $\theta(\cdot)$ is an s-vector of smooth functions of the population mean β_0 such that $s \leq p = m$; that is, $r(\beta_0) = \theta(\beta_0) - \theta^0$ (see Section 4.2). DiCiccio et al. (1991) show that the EL version of the criterion-based statistic $\mathcal{LR}_n^r = 2n\left(\hat{P}_n^{\rho}\left(\hat{\beta}^r, \hat{\lambda}^r\right) - \hat{P}_n^{\rho}\left(\hat{\beta}, \hat{\lambda}\right)\right)$ (Equation 12) is Bartlett correctable.[10] For the same setup, Baggerly (1998) proves that only EL is Bartlett correctable in the CR class of criteria. Chen & Cui (2007) provide the generalization to the overidentified moment condition setting $\mathbb{E}[g(z, \beta_0)] = 0$ (Equation 1), where $m \geq p$, dealing with the respective null and alternative hypotheses $H_0 : \beta_0 = \beta^0$ and $H_1 : \beta_0 \neq \beta^0$, demonstrating that the EL criterion statistic in Equation 12 is also Bartlett correctable in this case.[11] In cases in which interest solely concerns the null hypothesis $H_0 : \beta_{10} = \beta_1^0$ expressed in terms of a subvector of $\beta = \left(\beta_1', \beta_2'\right)'$, Chen & Cui (2006) show that, if the nuisance parameter vector β_2 is first profiled out from the EL criterion, then the resultant EL statistic for $H_0 : \beta_{10} = \beta_1^0$ against $H_1 : \beta_{10} \neq \beta_1^0$ is Bartlett correctable.

[9] Briefly, the Bartlett correction of the classical LR statistic $\mathcal{LR}$ takes the form $\mathcal{LR}_c = \mathcal{LR}/(1 + B_c/n)$, where the factor B_c is a function of moments of derivatives of the log-likelihood function that typically can be consistently estimated.

[10] Jing & Wood (1996) demonstrate that the EL(ET) criterion test statistic is not Bartlett correctable.

[11] Liu & Chen (2010) show that the test based on the adjusted EL criterion (see Section 5) achieves the same accuracy as the Bartlett-corrected EL statistic if the adjustment factor is set as $a_n = B_c/2$, where B_c denotes the Bartlett correction.

Matsushita & Otsu (2013) prove that the EL criterion function statistic in Equation 10 for overidentifying moment conditions is Bartlett correctable (see Section 4.1). Moreover, the adjusted EL criterion statistic (Chen et al. 2008, Liu & Chen 2010) (see Section 5) for testing overidentifying moments achieves the same accuracy as the Bartlett-corrected EL criterion statistic if the adjustment factor a_n is chosen such that $a_n = B_c/2$, where B_c is the Bartlett correction.

6.3. Large Deviations

Consider two distinct simple hypotheses $H_0 : \beta_0 = \beta^0$ and $H_1 : \beta_0 = \beta^1$, where β^0 and β^1 are known p-vectors of constants and $\beta^0 \neq \beta^1$. Let δ_0 and δ_1 denote the respective probabilities of type I and type II errors. Ideally, these probabilities should be set as low as possible. Because these probabilities are inversely related, Neyman-Pearson theory advises minimizing δ_1 (or equivalently maximizing power $1 - \delta_1$) for fixed δ_0. As is well known by the Neyman-Pearson lemma for a given size δ_0, the LR test of $H_0 : \beta_0 = \beta^0$ against $H_1 : \beta_0 = \beta^1$ is the most powerful.

Hoeffding (1965) considers a similar setting but with both type I and type II error probabilities, δ_{0n} and δ_{1n}, respectively, depending on the sample size n and approaching zero exponentially as n increases. Thus, the type I and II errors correspond to extreme tail events for sufficiently large n. Hoeffding (1965) shows the large deviation result that, among those tests of $H_0 : \beta_0 = \beta^0$ against $H_1 : \beta_0 = \beta^1$ that satisfy

$$\limsup_{n\to\infty} n^{-1} \log \delta_{0n} \leq -\eta \tag{17}$$

for η fixed, the LR test minimizes

$$\limsup_{n\to\infty} n^{-1} \log \delta_{1n}; \tag{18}$$

that is, the LR test minimizes the probability of a type II error.

Kitamura et al. (2012) apply the large deviation theory of Hoeffding (1965) to tests of overidentifying moment restrictions (see Section 4.1). In general, there is no test that satisfies Equation 17. However, if certain probability distributions that satisfy the moment restrictions in Equation 1 are eliminated, then the following results hold: (*a*) The test of overidentifying moment restrictions $H_0 : \mathbb{E}[g(z, \beta)] = 0$ for some $\beta \in \mathcal{B}$ formed from the EL-based criterion $\mathcal{LR}_n = 2n\hat{P}_n^{\rho}(\hat{\beta}, \hat{\lambda})$ (Equation 10) satisfies Equation 17; (*b*) among all tests that satisfy Equation 17, the EL-based criterion test minimizes Equation 18 if the alternative hypothesis $H_1 : \mathbb{E}[g(z, \beta)] \neq 0$ for all $\beta \in \mathcal{B}$ deviates sufficiently from H_0.

6.4. Robustness

The robustness properties of the ML estimator are examined by Beran (1977). Small deviations from the assumed parametric density function can lead to large variations in the log-likelihood function, demonstrating a lack of robustness of ML in this sense. Beran (1977) shows that an alternative parametric estimator based on the minimization of a discrepancy measure formulated in terms of Hellinger distance is robust but also shares the asymptotic efficiency property of ML.

Let $f_0(\cdot)$ denote the density function of the data observation z commensurate with the moment condition $\mathbb{E}[g(z, \beta)] = 0$ (Equation 1) satisfied at the unique value β_0 of $\beta \in \mathcal{B}$. Based on random samples drawn from density functions in a neighborhood of $f_0(\cdot)$, Kitamura et al. (2013) analyze the robustness properties of estimators of a known scalar function $m(\cdot)$ of β_0 in terms of their asymptotic maximum bias and mean squared error. In the class of Fisher consistent and regular estimators, which includes GMM and GEL, maximum bias is minimized by the minimum

Hellinger distance estimator (MHDE) computed using the trimmed moment indicators $g(z,\beta)I(\sup_{\beta\in\mathcal{B}}||g(z,\beta)|| \le m_n)$, where $\{m_n\}$ is a positive-valued sequence that approaches infinity with sample size n. Additionally, the mean squared error is also minimized by MHDE based on the moment indicator vector $g(z, \beta)$. MHDE corresponds to the CR discrepancy measure with parameter $\tau = -1/2$; thus, MHDE is also GEL and is thereby asymptotically efficient (see Section 3).

7. CONDITIONAL MOMENTS

Many empirical problems concern models defined through conditional rather than unconditional moment restrictions. Let $u(z, \beta)$ be a known J-vector of functions of the random vector of observables z and the unknown p-vector of parameters β, which, as before, constitute the object of inferential interest. We consider models defined by the following conditional moments:

$$\mathbb{E}\big[u(z,\beta_0)|x\big] = 0, \tag{19}$$

where x is a d_x-dimensional subvector of z.

The following examples illustrate this framework.

> **Example 6 (quantile regression continued):** The quantile regression model is redefined by the conditional probability statement $\mathcal{P}\{y \le x'\beta_0|x\} = \theta$. The conditional moment condition in Equation 19 defining θ-conditional quantile regression is $\mathbb{E}[\theta - I(y \le x'\beta_0)|x] = 0$ or
>
> $$\mathbb{E}\big[\theta - I(u \le 0)|x\big] = 0,$$
>
> where $u = y - x'\beta_0$. Here β_0 is the unique value of β that satisfies $\mathbb{E}[\theta - I(y \le x'\beta)|x] = 0$. Similar to that described in Section 2, the unconditional moment restriction $\mathbb{E}[x(\theta - I(u \le 0))] = 0$ holds as does $\mathbb{E}[q(x)(\theta - I(u \le 0))] = 0$ for suitably defined vectors of functions $q(\cdot)$ of x.
>
> **Example 7 (instrumental variables continued):** The conditional mean restriction $\mathbb{E}[y - x'\beta_0|w] = 0$ or $\mathbb{E}[u|w] = 0$ (see Equation 19), where $u = y - x'\beta_0$, is often assumed in a linear regression setting (see, e.g., Davidson & MacKinnon 2004, Greene 2008). Here β_0 is the unique value of β that satisfies $\mathbb{E}[y - x'\beta|w] = 0$. Similar to Example 6 above, the unconditional moment restriction $\mathbb{E}[q(w)u] = 0$ is implied, where $q(\cdot)$ is a vector of functions of the instruments w; in particular, $\mathbb{E}[wu] = 0$ holds. For standard linear regression, the conditional mean restriction $\mathbb{E}[y|x] = x'\beta_0$ or $\mathbb{E}[u|x] = 0$ would be the standard assumption. It is well known (see Cragg 1983) that unless the linear regression model is conditionally homoscedastic (i.e., var$[u|x]$ is constant or invariant to x), instrumental variable estimation of β_0 based on the unconditional moment restriction $\mathbb{E}[q(x)u] = 0$ when $q(x)$ includes x is more efficient than least squares.

GMM- and EL-type estimators of β_0 that achieve the semiparametric efficiency lower bound (Chamberlain 1987) are proposed by Donald et al. (2003), Kitamura et al. (2004), and Zhang & Gijbels (2003). Donald et al. (2003) use unconditional moment restrictions based on particular classes of approximating functions, such as splines and power series, whereas Kitamura et al. (2004) and Zhang & Gijbels (2003) employ kernel smoothed moment indicator functions. Otsu (2007) extends the conditional EL approach of Kitamura et al. (2004) and Zhang & Gijbels (2003) to conditional moment restriction models that incorporate unknown infinite-dimensional functions

requiring nonparametric estimation, models studied previously by Ai & Chen (2003) using the method of sieves applied to GMM.

7.1. Approximating Functions

Donald et al. (2003) note that any function can be approximated arbitrarily well by linear combinations of certain basis or approximating functions when the number of functions is allowed to approach infinity with sample size n. Particular examples of admissible classes of approximating functions are splines, power series, and Fourier series (see, e.g., Powell 1981).

More specifically, let K be the number of approximating functions and define $q^K(x) = (q_{1K}(x), \ldots, q_{KK}(x))'$ as the K-vector of approximating functions. The consequent vector of unconditional moment indicators is given by (see Equation 1)

$$g^K(z, \beta) = u(z, \beta) \otimes q^K(x). \tag{20}$$

Donald et al. (2003, lemma 2.1, p. 58) demonstrate a formal equivalence between the sequence of unconditional moment constraints $\mathbb{E}[g^K(z, \beta_0)] = 0$ (Equation 20), $K \to \infty$, and the conditional moment restriction $\mathbb{E}[u(z, \beta_0)|x] = 0$ (Equation 19). More precisely, by the law of iterated expectations, one obtains $\mathbb{E}[g^K(z, \beta_0)] = 0$ for all K if $\mathbb{E}[u(z, \beta_0)|x] = 0$ and, moreover, $\mathbb{E}[g^K(z, \beta_0)] \neq 0$ for all K large enough if $\mathbb{E}[u(z, \beta_0)|x] \neq 0$.

Consequently, EL, GMM, and GEL may be applied using the unconditional moment indicator vector $g^K(z, \beta)$ (Equation 20). Donald et al. (2003) show that if K approaches infinity at an appropriate rate dependent on the approximating functions and the estimator employed, the resultant estimators are consistent and achieve the semiparametric efficiency lower bound $\mathbb{E}[D(x)'\Sigma(x)^{-1}D(x)]^{-1}$ (Chamberlain 1987), where $D(x) = \mathbb{E}[\partial u(z, \beta_0)/\partial \beta'|x]$ and $\Sigma(x) = \mathbb{E}[u(z, \beta_0)u(z, \beta_0)'|x]$.

7.2. Conditional (G)EL

Let $u_i(\beta) = u(z_i, \beta)$, $i = 1, \ldots, n$. Kitamura et al. (2004) (see also Zhang & Gijbels 2003) modify EL by scaling the standard EL criterion in Equation 5 using the positive weights w_{ij}, $i, j = 1, \ldots, n$; that is,

$$\mathcal{CEL}_n(\beta, \lambda) = \sum_{i=1}^{n} \mathbb{T}_{i,n} \sum_{j=1}^{n} w_{ij} \log\left(1 + \lambda_i' u_j(\beta)\right) \Big/ n, \tag{21}$$

where $\lambda = \left(\lambda_1', \ldots, \lambda_n'\right)'$, $w_{ij} = \mathcal{K}_{ij} / \sum_{k=1}^{n} \mathcal{K}_{ik}$, $\mathcal{K}_{ij} = \mathcal{K}\left((x_i - x_j)/b_n\right)$, $\mathcal{K}(\cdot)$ is a symmetric positive kernel, and b_n is a bandwidth parameter. The trimming function $\mathbb{T}_{i,n}$ is required to ensure that the denominator of the weights w_{ij} is bounded away from zero; that is, $\mathbb{T}_{i,n} = I\left(\hat{h}(x_i) \geq b_n^{\tau}\right)$ for some $\tau \in (0, 1)$, where $\hat{h}(x) = \sum_{j=1}^{n} \mathcal{K}\left((x - x_i)/b_n\right)/n b_n^{d_x}$ is the standard kernel estimator for the density $h(\cdot)$ of x and $I(\cdot)$ is an indicator function.[12] Note that $\mathcal{CEL}_n(\beta, \lambda)$ employs the Nadaraya-Watson estimator $\sum_{j=1}^{n} w_{ij} \log\left(1 + \lambda_i' u_j(\beta)\right)$ of the conditional expectation of

[12]The criterion $\mathcal{CEL}_n(\beta, \lambda)$ (Equation 21) is obtained from the program $\max_{\beta, \{\pi_{ij}\}_{i,j=1}^{n}} \sum_{i=1}^{n} \mathbb{T}_{i,n} \sum_{j=1}^{n} w_{ij} \log \pi_{ij}$ subject to $\pi_{ij} \geq 0$, $\sum_{j=1}^{n} \pi_{ij} = 1$, and $\sum_{j=1}^{n} \pi_{ij} u_j(\beta) = 1$, $i, j = 1, \ldots, n$. Note that π_{ij} has the interpretation as the probability $\mathcal{P}\{z = z_j | x = x_i\}$, $i, j = 1, \ldots, n$, and λ_i is the Lagrange multiplier associated with the sample moment constraint $\sum_{j=1}^{n} \pi_{ij} u_j(\beta) = 1$, $i = 1, \ldots, n$.

$\log\left(1+\lambda_i' u_j(\beta)\right)$ given x_i (i.e., $\mathbb{E}\left[\log\left(1+\lambda_i' u_j(\beta)\right)\middle| x_i\right], i=1,\ldots,n$) and thus may be regarded as an estimator of the average conditional expectation $\sum_{i=1}^n \mathbb{E}\left[\log\left(1+\lambda_i' u_i(\beta)\right)\middle| x_i\right]\Big/ n$.

Let $\Lambda_n = \left\{\lambda \in \mathbb{R}^J : ||\lambda|| \leq Cn^{-1/m}\right\}$ for some finite constant $C > 0$.[13] The conditional EL estimator is the solution to a saddle point problem

$$\hat{\beta} = \arg\inf_{\beta\in\mathcal{B}} \sum_{i=1}^n \mathbb{T}_{i,n} \sup_{\lambda_i\in\Lambda_n} \sum_{j=1}^n w_{ij} \log\left(1+\lambda_i' u_j(\beta)\right)\Big/ n \tag{22}$$

with the Lagrange multiplier estimator $\hat{\lambda}_i(\beta)$ defined by $\hat{\lambda}_i(\beta) = \arg\max_{\lambda_i\in\Lambda_n} \sum_{j=1}^n w_{ij} \ln\left(1+\lambda_i' u(z_j,\beta)\right)$. The conditional EL estimator $\hat{\beta}$ (Equation 22) is consistent and achieves the semiparametric efficiency lower bound $\mathbb{E}[D(x)'\Sigma(x)^{-1}D(x)]^{-1}$ (see Kitamura et al. 2004).[14]

Conditional EL was subsequently generalized in Smith (2007a,b) for GEL and the CR power divergence family with criterion $\sum_{i=1}^n \mathbb{T}_{i,n} \sum_{j=1}^n w_{ij}\left[\rho\left(\lambda_i' u_j(\beta)\right) - \rho(0)\right]\Big/ n$, where $\rho(\cdot)$ is defined in Section 3.4 (see also Antoine et al. 2007, which proposes a similar conditional estimator based on CUE).

7.3. Unknown Functions

Consider the generalized form of the vector of conditional moment restrictions $\mathbb{E}[u(z,\beta)|x] = 0$ (Equation 19) given by

$$\mathbb{E}\left[u\left(z,\beta_0,h_0(x_z)\right)\middle| x\right] = 0. \tag{23}$$

Here, as above, $u(z, \beta_0, h_0(x_z))$ is a J-vector of known functions but now includes the unknown vector $h_0(\cdot)$ of smooth functions of the subvector x_z of the conditioning variables x as an argument. Although β_0 remains of central inferential interest, the unknown function $h_0(\cdot)$ is of interest too. Examples of this general framework include partially linear regression $u\left(z,\beta_0,h_0(x_z)\right) = y - x_1'\beta_0 - h_0(x_z)$, where $x = (x_1', x_z')'$ (Robinson 1988a), and single index regression $u\left(z,\beta_0, h_0(x_z)\right) = y - h_0(x_z'\beta_0)$, where $x = x_z$ (Powell et al. 1989, Ichimura 1993).

Let the true parameter vector $\alpha_0 = \left(\beta_0', h_0'\right)'$ with parameter space $\mathcal{A} = \mathcal{B}\times\mathcal{H}$. Consequently, the conditional moment restriction in Equation 23 may be rewritten as $\mathbb{E}[u(z, \alpha_0)|x] = 0$. Although with this redefinition the conditional moment restriction in Equation 23 now superficially resembles Equation 19, Kitamura et al.'s (2004) conditional EL estimator clearly cannot be applied without modification, as α_0 contains the infinite-dimensional parameter $h_0(\cdot)$.

Let $u_i(\alpha) = u(z_i, \alpha)$, $i = 1,\ldots,n$. The penalized EL criterion proposed by Otsu (2007) adopts Shen's (1997) approach, modifying the conditional EL criterion $\mathcal{CEL}_n(\beta,\lambda)$ (Equation 21) as

$$\mathcal{PEL}_n(\alpha,\lambda) = \sum_{i=1}^n \mathbb{T}_{i,n} \sum_{j=1}^n w_{ij} \log\left(1+\lambda' u_j(\alpha)\right)\Big/ n - \phi_n \mathcal{J}(h),$$

with the incorporation of the penalty function $\mathcal{J}(\cdot)$ to impose some restrictions on the parameter space $\mathcal{A}$; the positive-valued sequence $\{\phi_n\}$ of penalization constants is chosen so as to converge to

[13] For technical reasons, m is a positive integer such that $m \geq 8$.

[14] Tripathi & Kitamura (2003) propose a test statistic based on conditional EL (see also Smith 2007a,b for conditional GEL-based test statistics).

zero with sample size n at rate $o(n^{-1/2})$. Examples of penalty functions $\mathcal{J}(\cdot)$ may be found in Shen (1997, section 3), for example, to impose twice differentiability on the resultant estimator of h_0. The penalized EL estimator $\hat{\alpha}$ is the solution to a saddle point problem

$$\hat{\alpha} = \arg \inf_{\alpha \in \mathcal{A}} \sum_{i=1}^{n} \mathbb{T}_{i,n} \sup_{\lambda_i \in \mathbb{R}^l} \sum_{j=1}^{n} w_{ij} \log\Big(1 + \lambda_i' u_j(\beta)\Big) \Big/ n, \tag{24}$$

with the Lagrange multiplier estimator $\hat{\lambda}_i(\beta)$ defined by $\hat{\lambda}_i(\beta) = \arg\max_{\lambda_i \in \mathbb{R}^l} \sum_{j=1}^{n} w_{ij} \log(1 + \lambda_i' u(z_j, \beta))$. Otsu (2007) proves the consistency of the penalized EL estimator $\hat{\alpha} = \big(\hat{\beta}', \hat{h}'\big)'$ for $\alpha_0 = \big(\beta_0', h_0'\big)'$ together with the respective convergence rates of $\hat{\beta}$ and $\hat{h}$. Moreover, the penalized EL estimator $\hat{\beta}$ (Equation 24) of β_0 is asymptotically normal and achieves the semiparametric efficiency lower bound $\mathbb{E}[D(x)'\Sigma(x)^{-1}D(x)]^{-1}$, where now $D(x) = \mathbb{E}[\partial u(z, \alpha_0)/\partial \beta' | x]$ and $\Sigma(x) = \mathbb{E}[u(z, \alpha_0)u(z, \alpha_0)' | x]$.

In an earlier paper, Ai & Chen (2003) suggest a sieve minimum distance approach similar to GMM in which the conditional moment indicator vector $u_i(\cdot)$ is estimated using sieves rather than kernel functions and the unknown functions comprising h_0 are also likewise approximated. Unlike Otsu (2007), Ai & Chen (2003) assume $\mathcal{A}$ is compact, which has the advantage of allowing the unknown vector of functions h_0 to depend on z rather than solely a subvector of x, thus permitting the inclusion of endogenous variables. More recently, for a similar setup permitting endogenous components of the unknown function vector h_0, Otsu (2011) applies Kitamura et al.'s (2004) conditional EL method with h_0 approximated by sieves as in Ai & Chen (2003). Chen & Pouzo (2009) generalize Ai & Chen's (2003) sieve minimum distance method to allow for nonsmooth functions to comprise h_0 together with a bootstrap procedure for improved inference.[15] All these methods result in consistent and asymptotically equivalent normally distributed estimators of β_0 that achieve the semiparametric efficiency lower bound $\mathbb{E}[D(x)'\Sigma(x)^{-1}D(x)]^{-1}$.

8. WEAKLY DEPENDENT DATA

In this section, z_t, $t = 1, \ldots, T$, denotes T observations on a finite-dimensional stationary and strongly mixing process $\{z_t\}_{t=1}^{\infty}$. The moment indicator vector $g(z_t, \beta)$ is defined, as previously, as an m-vector of known functions of the data observation z_t and the p-vector β of unknown parameters that are the object of inferential interest, where $m \geq p$. It is assumed that the true parameter vector β_0 uniquely satisfies the moment condition

$$\mathbb{E}\big[g(z_t, \beta)\big] = 0, \tag{25}$$

where $\mathbb{E}[\cdot]$ denotes expectation taken with respect to the unknown distribution of z_t. Because Equation 25 may arise in many cases from conditional moment restrictions, z_t may also include lagged endogenous and current and lagged values of exogenous variables.

Define $g_t(\beta) = g(z_t, \beta), t = 1, \ldots, T$, and $\hat{g}(\beta) = T^{-1}\sum_{t=1}^{T} g_t(\beta)$. Let $k(\cdot)$ denote a kernel function that satisfies the mild regularity conditions stated in Smith (2011) and define $k_j = \int_{-\infty}^{\infty} k(a)^j da$, $j = 1, 2, 3$, with $k = k_1/k_2$. The bandwidth parameter S_T diverges to infinity at an appropriate rate dependent on the kernel function $k(\cdot)$ and sample size T.

[15]For stationary and ergodic data, Chen & Pouzo (2012) establish convergence rates for a penalized sieve minimum distance estimator of h_0 in circumstances that similarly permit nonsmooth unknown functions and the possible inclusion of endogenous variables. The penalization (see Otsu 2007) avoids the necessity of restricting the parameter space $\mathcal{A}$ to be compact and may ease computational difficulties associated with the sieve minimum distance methods of Ai & Chen (2003) and Chen & Pouzo (2009).

8.1. Efficient (G)EL

Kitamura & Stutzer (1997) observed that applying standard ET to the moment indicators $\{g_t(\beta)\}_{t=1}^T$ results in a consistent but asymptotically inefficient estimator of β_0 if there is dependence. To deal with this problem, they modify the ET criterion by basing it on a smoothed version of the moment indicators $\{g_t(\beta)\}_{t=1}^T$ obtained using the truncated or uniform kernel function. Smith (1997, 2011) discusses GEL employing general kernel functions. For suitable choices of the kernel function $k(\cdot)$, GEL is asymptotically efficient.[16] Kitamura (1997) suggests an alternative approach using blockwise EL. The exposition that follows is based on the approach in Smith (2011).

Define the smoothed moment indicators

$$g_{tT}(\beta) = \frac{1}{S_T}\sum_{s=t-T}^{t-1} k\left(\frac{s}{S_T}\right) g_{t-s}(\beta),\ t = 1,\ldots,T. \tag{26}$$

Examples of admissible kernel functions $k(\cdot)$ are the truncated or uniform kernel $k_{\mathrm{TR}}(x) = I(|x| \le 1)$ and the Bartlett kernel $k_{\mathrm{BT}}(x) = (1 - |x|)I(|x| \le 1)$.

GEL criteria appropriate for weakly dependent data are defined by (see Section 3.4)

$$\hat{P}_T^{\rho}(\beta,\lambda) = \sum\nolimits_{t=1}^{T}\left[\rho\left(k\lambda' g_{tT}(\beta)\right) - \rho(0)\right]\Big/T,$$

with the GEL estimator then given by

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} \sup_{\lambda\in\Lambda_T} \hat{P}_T^{\rho}(\beta,\lambda),$$

where $\Lambda_T = \{\lambda : ||\lambda|| \le C_T\}$, with C_T a positive sequence that depends on T and converges to zero at an appropriate rate [see Smith 2011, assumption 2.4(b), p. 1200]. Let $\hat{\lambda}(\beta) = \arg\sup_{\lambda\in\Lambda_T} \hat{P}_T^{\rho}(\beta,\lambda)$ with $\hat{\lambda} = \hat{\lambda}\left(\hat{\beta}\right)$.

Define $\Sigma = (G'\Omega^{-1}G)^{-1}$ and $P = \Omega^{-1} - \Omega^{-1}G\Sigma G'\Omega^{-1}$. Then, under standard regularity conditions,

$$T^{1/2}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0,\Sigma),\quad \left(T/S_T^2\right)^{1/2}\hat{\lambda} \xrightarrow{d} N(0,P),$$

and the GEL estimator $\hat{\beta}$ and the auxiliary parameter estimator $\hat{\lambda}$ are asymptotically uncorrelated. Consequently, GEL is asymptotically equivalent to asymptotically efficient GMM.

8.2. Higher-Order Properties

The literature on the higher-order properties of (G)EL for time series data is relatively limited. Kitamura (1997) shows the Bartlett correctability of the blockwise EL criterion statistic in the time series context for smooth functions of the mean (see Section 6.2). Anatolyev (2005) investigates the asymptotic bias of 2SGMM and GEL estimators based on the smoothed moment indicators in Equation 26.

[16] In general, the first-order condition for the GEL estimator $\tilde{\beta}$ using the unsmoothed moment indicators $\{g_t(\beta)\}_{t=1}^T$ may be expressed as $\left[\sum_{t=1}^T \tilde{\pi}_t G_t(\tilde{\beta})\right]'\left[\sum_{t=1}^T \tilde{p}_t g_t(\tilde{\beta}) g_t(\tilde{\beta})'\right]^{-1}\hat{g}(\tilde{\beta}) = 0$, where $G_t(\beta) = \partial g_t(\beta)/\partial\beta'$, $\tilde{\pi}_t = \rho_1\left(\tilde{\lambda}' g_t(\tilde{\beta})\right)\Big/\sum_{s=1}^T \rho_1\left(\tilde{\lambda}' g_s(\tilde{\beta})\right)$, $\tilde{p}_t = p\left(\tilde{\lambda}' g_t(\tilde{\beta})\right)\Big/\sum_{s=1}^T p\left(\tilde{\lambda}' g_s(\tilde{\beta})\right)$, $t = 1,\ldots,T$, with the function $p(\cdot)$ defined as $p(v) = [\rho_1(v) + 1]/v$, $v \ne 0$, $p(0) = -1$, and $\tilde{\lambda} = \sup_{\lambda\in\Lambda_T} \sum_{t=1}^T \left[\rho\left(\lambda' g_t(\tilde{\beta})\right) - \rho(0)\right]\Big/T$ (see Equation 8, with Λ_T defined below). Although $\sum_{t=1}^T \tilde{\pi}_t G_t(\tilde{\beta})$ is a consistent estimator for G, $\sum_{t=1}^T \tilde{p}_t g_t(\tilde{\beta}) g_t(\tilde{\beta})'$ consistently estimates the short-run variance matrix $\mathbb{E}[g_t(\beta_0) g_t(\beta_0)']$ rather than the long-run variance matrix Ω required for asymptotic efficiency.

Anatolyev (2005) considers the 2SGMM estimator with the weighting matrix the inverse of the HAC (heteroscedastic autocorrelation consistent) estimator $S_T\sum_{t=1}^{T} g_{tT}(\tilde{\beta})g_{tT}(\tilde{\beta})'/T$ (Smith 2005) of the moment indicator vector long-run variance matrix Ω, where $\tilde{\beta}$ is a preliminary root-T consistent estimator of β_0; CUE with the same form of weighting matrix $\left[S_T\sum_{t=1}^{T} g_{tT}(\beta)g_{tT}(\beta)'/T\right]^{-1}$; and GEL as in Section 8.1. Anatolyev (2005) confines consideration to kernels $k(\cdot)$ with bounded support (i.e., nonzero on an open interval including zero and zero elsewhere). Similarly to the discussion in Section 6.1, the asymptotic bias of GMM and GEL may be decomposed into a number of factors, each of which has the same interpretation as in Section 6.1; that is, for GMM,[17]

$$\text{abias}\left[\hat{\beta}\right] = B_I + B_G + B_\Omega + B_W,$$

and for GEL,

$$\text{abias}\left[\hat{\beta}\right] = B_I + \left(1 + \rho_3 k_1 k_3/2k_2^2\right)B_\Omega.$$

Note that, because $\rho_3 = 0$, the CUE asymptotic bias is $B_I + B_\Omega$ (see Section 6.1). For further details, readers are referred to Anatolyev (2005, theorem 1, p. 988).

The interpretation of these asymptotic bias terms is similar to those given above for the cross-sectional setting, although their mathematical expressions differ. In particular, as before, B_I coincides with the asymptotic bias for an infeasible GMM estimator obtained from the first-order conditions $G'\Omega^{-1}\hat{g}(\beta) = 0$, B_G results from the estimation of G and is absent for GEL, B_Ω is for (implicit) estimation of Ω, and B_W is from the preliminary consistent estimation of β_0 in 2SGMM.

Unlike cross-sectional data, the B_Ω term does not vanish for EL unless $k_2^2 = k_1 k_3$, which occurs for the truncated or uniform kernel considered by Kitamura & Stutzer (1997). More generally, this term is not present for GEL criteria $\rho(\cdot)$ and kernel functions $k(\cdot)$ such that $\rho_3 = -2(k_2^2)/(k_1 k_3)$. For the GEL class $\rho(v) = -(1 + \tau v)^{(1+\tau)/\tau}/(1 + \tau)$ equivalent to the CR power divergence family, the choice $\tau = \left(k_1 k_3 - 2\left(k_2^2\right)\right)\Big/(k_1 k_3)$ is required.

9. CONCLUSIONS

Several open research areas remain. For example, as noted in Section 5, although (G)EL methods have attractive theoretical large sample properties, their computation raises serious practical difficulties because of the induced nonlinearity arising from the necessity to solve a saddle point problem. Fast and reliable algorithms for (G)EL when the number of moment restrictions and parameters to be estimated are large would enable (G)EL to be applied to a wider variety of empirical problems than is currently the case and would allow for more substantial and detailed simulation studies of these techniques to be undertaken. Bootstrap methods specifically designed for the application of (G)EL are scarce, although a notable exception is provided by Canay (2010). As noted in Section 3, (G)EL imposes the moment

[17]Let a be the $m \times 1$ vector such that $a_j = tr\left(\Sigma\mathbb{E}\left[\partial^2 g_t^j(\beta_0)/\partial\beta\partial\beta'\right]\right)/2, j = 1, \ldots, m$, where $g_t^j(\beta)$ denotes the j-th element of $g_t(\beta)$, $g_t = g_t(\beta_0)$, $G_t(\beta) = \partial g_t(\beta)/\partial\beta'$, and $G_t = G_t(\beta_0)$. Also let $H_W = (G'W^{-1}G)^{-1}G'W^{-1}$, $H = \Sigma G'\Omega^{-1}$, $\overline{\Omega}_{\beta_j}(u) = \mathbb{E}\left[\partial(g_t(\beta_0)g_{t-u}(\beta_0)')/\partial\beta_j\right]$, and e_j the j-th unit vector. Then $B_I = H\left(-a + \sum_{u=-\infty}^{\infty}\mathbb{E}[G_t H g_{t-u}]\right)/T$, $B_G = -\Sigma\sum_{u=-\infty}^{\infty}\mathbb{E}\left[G_t' P g_{t-u}\right]/T$, $B_\Omega = H\sum_{u,v=-\infty}^{\infty}\mathbb{E}\left[g_t g_{t-u}' P g_{t-v}\right]/T$, and $B_W = -H\sum_{j=1}^{p}\sum_{u=-\infty}^{\infty}\overline{\Omega}_{\beta_j}(u)(H_W - H)'e_j/T$.

restrictions in the sample. Brown & Newey (2002) exploit this feature for the cross-sectional setting by reweighting moment indicator observations using the empirical probabilities. Therefore, unlike Hall & Horowitz's (1996) method, their bootstrap method based on the resampling of the reweighted moment indicators does not require moments to be explicitly centered. In general, the higher-order properties of this and other related procedures remain to be investigated.

The literature on estimation and inference for models specified by moment condition constraints is vast. Hence, because of space limitations, this review can only be partial in terms of its coverage. Several other important areas are also currently exciting considerable research effort.

This review concentrates above on models in which parameters are point identified. Moment condition models in which parameters are only set or partially identified have received a great deal of attention in the recent literature (see, e.g., Andrews & Shi 2013, which considers moment inequality restrictions, and references therein).

The extensive literature on weak identification, in particular, weak instruments in the regression context, was initiated by Angrist & Krueger (1991), who study the returns to education. Weak identification essentially concerns the lack of correlation between the moment indicator vector and the (implicit) score vector associated with the true model. Standard inferential tools such as LR and Wald test statistics no longer have the standard limiting normal or chi-squared distributions. Several methods have been proposed to ameliorate this problem, primarily related to score or Lagrange multiplier statistics, as discussed in Section 4 (see, e.g., Kleibergen 2005, Otsu 2006, Guggenberger et al. 2012). Newey & Windmeijer (2009) obtain the limiting properties of GMM and (G)EL when there are many weak moments and, in particular, show that the respective variance matrices are inflated in comparison to the standard variance matrix expression given in Section 3 for efficient 2SGMM and GEL.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Ai C, Chen X. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71:1795–843

Ali SM, Silvey SD. 1966. A general class of coefficient of divergence of one distribution from another. *J. R. Stat. Soc. B* 28:131–42

Amemiya T. 1974. The nonlinear two-stage least-squares estimator. *J. Econom.* 2:105–10

Anatolyev S. 2005. GMM, GEL, serial correlation, and asymptotic bias. *Econometrica* 73:983–1002

Anatolyev S, Gospodinov N. 2011. *Methods for Estimation and Inference in Modern Econometrics*. Boca Raton, FL: Chapman & Hall/CRC

Andrews DWK, Shi X. 2013. Inference based on conditional moment inequalities. *Econometrica* 81:609–66

Andrews DWK, Stock JH, eds. 2005. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge, UK: Cambridge Univ. Press

Angrist JD, Krueger AB. 1991. Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* 106:979–1014

Antoine B, Bonnal H, Renault E. 2007. On the efficient use of the informational content of estimating equations: implied probabilities and Euclidean empirical likelihood. *J. Econom.* 138:461–87

Back K, Brown DP. 1993. Implied probabilities in GMM estimators. *Econometrica* 61:971–75

Baggerly K. 1998. Empirical likelihood as a goodness-of-fit measure. *Biometrika* 85:535–47

Bartlett MS. 1937. Properties of sufficiency and statistical tests. *Proc. R. Soc. A* 160:268–82
Beran R. 1977. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* 5:445–63
Brown BW, Newey WK. 1998. Efficient semiparametric estimation of expectations. *Econometrica* 66:453–64
Brown BW, Newey WK. 2002. Generalized method of moments, efficient bootstrapping, and improved inference. *J. Bus. Econ. Stat.* 20:507–17
Canay IA. 2010. EL inference for partially identified models: large deviations optimality and bootstrap validity. *J. Econom.* 156:408–25
Chamberlain G. 1987. Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econom.* 34:305–34
Chaussé P. 2010. Computing generalized method of moments and generalized empirical likelihood with R. *J. Stat. Softw.* 34(11):1–35
Chen J, Variyath AM, Abraham B. 2008. Adjusted empirical likelihood and its properties. *J. Comput. Graph. Stat.* 17:426–43
Chen SX, Cui H-J. 2006. On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* 93:215–20
Chen SX, Cui H-J. 2007. On the second order properties of empirical likelihood with moment restrictions. *J. Econom.* 141:492–516
Chen X, Pouzo D. 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econom.* 152:46–60
Chen X, Pouzo D. 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* 80:277–321
Chesher A, Smith RJ. 1997. Likelihood ratio specification tests. *Econometrica* 65:627–46
Corcoran S. 1998. Bartlett adjustment of empirical discrepancy statistics. *Biometrika* 85:965–72
Cragg JG. 1983. More efficient estimation in the presence of heteroscedasticity of unknown form. *Econometrica* 51:751–63
Cressie N, Read T. 1984. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. B* 46:440–64
Cribari-Neto F, Cordeiro GM. 1996. On Bartlett and Bartlett-type corrections. *Econom. Rev.* 15:339–67
Csiszar I. 1963. Eine informations theoretische ungleichungen und ihre anwendung auf den beweis der ergodicitat von Markoffschen ketten. *Publ. Math. Inst. Hung. Acad. Sci.* 8:85–108
Davidson R, MacKinnon JG. 2004. *Econometric Theory and Methods*. New York: Oxford Univ. Press
DiCiccio T, Hall P, Romano J. 1991. Empirical likelihood is Bartlett-correctable. *Ann. Stat.* 19:1053–61
Donald SG, Imbens GW, Newey WK. 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *J. Econom.* 117:55–93
Fan Y, Gentry M, Li T. 2011. A new class of asymptotically efficient estimators for moment condition models. *J. Econom.* 162:268–77
Ghosh JK. 1994. *Higher Order Asymptotics*. NSF-CBMS Reg. Conf. Ser. Probab. Stat. 4. Hayward, CA: Inst. Math. Stat.
Goldberger AS. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard Univ. Press
Greene WH. 2008. *Econometric Analysis*. Upper Saddle River, NJ: Pearson Prentice Hall. 6th ed.
Guggenberger P. 2008. Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econom. Rev.* 27:526–41
Guggenberger P, Ramalho JJS, Smith RJ. 2012. GEL statistics under weak identification. *J. Econom.* 170:331–49
Hall P, Horowitz JL. 1996. Bootstrap critical values for tests based on generalized-method-of-moment estimators. *Econometrica* 64:891–916
Hansen LP. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54
Hansen LP, Heaton J, Yaron A. 1996. Finite-sample properties of some alternative GMM estimators. *J. Bus. Econ. Stat.* 14:262–80
Hjört NL, McKeague IW, Van Keilegom I. 2009. Extending the scope of empirical likelihood. *Ann. Stat.* 37:1079–111

Hoeffding W. 1965. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* 36:369–408
Ichimura H. 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *J. Econom.* 58:71–120
Imbens GW. 1997. One-step estimators for over-identified generalized method of moments models. *Rev. Econ. Stud.* 64:359–83
Imbens GW. 2002. Generalized method of moments and empirical likelihood. *J. Bus. Econ. Stat.* 20:493–506
Imbens GW, Spady RH. 2002. Confidence intervals in generalized method of moments models. *J. Econom.* 107:87–98
Imbens GW, Spady RH. 2005. The performance of empirical likelihood and its generalizations. See Andrews & Stock 2005, pp. 216–44
Imbens GW, Spady RH, Johnson P. 1998. Information theoretic approaches to inference in moment condition models. *Econometrica* 66:333–57
Jing B-Y, Wood ATA. 1996. Exponential empirical likelihood is not Bartlett correctable. *Ann. Stat.* 24:365–69
Kitamura Y. 1997. Empirical likelihood methods with weakly dependent processes. *Ann. Stat.* 25:2084–102
Kitamura Y. 2001. Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* 69:1661–72
Kitamura Y. 2007. Empirical likelihood methods in econometrics: theory and practice. In *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. 3, ed. RW Blundell, WK Newey, T Persson, pp. 174–237. Cambridge, UK: Cambridge Univ. Press
Kitamura Y, Otsu T, Evdokimov K. 2013. Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica* 81:1185–201
Kitamura Y, Santos A, Shaikh AM. 2012. On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* 80:413–23
Kitamura Y, Stutzer M. 1997. An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65:861–74
Kitamura Y, Tripathi G, Ahn H. 2004. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica* 72:1667–714
Kleibergen FR. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73:1103–23
Liu Y, Chen J. 2010. Adjusted empirical likelihood with high-order precision. *Ann. Stat.* 38:1341–62
Manski CF. 1988. *Analog Estimation Methods in Econometrics*. New York: Chapman & Hall
Matsushita Y, Otsu T. 2013. Second-order refinement of empirical likelihood for testing overidentifying restrictions. *Econ. Theory* 29:324–53
Mittelhammer RC, Judge GG, Schoenberg R. 2005. Empirical evidence concerning the finite sample performance of EL-type structural equation estimation and inference methods. See Andrews & Stock 2005, pp. 282–305
Newey WK, Ramalho JJS, Smith RJ. 2005. Asymptotic bias for GMM and GEL estimators with estimated nuisance parameters. See Andrews & Stock 2005, pp. 245–81
Newey WK, Smith RJ. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72:219–55
Newey WK, West KD. 1987. Hypothesis testing with efficient method of moments estimation. *Int. Econ. Rev.* 28:777–87
Newey WK, Windmeijer F. 2009. Generalized method of moments with many weak moment conditions. *Econometrica* 77:687–719
Otsu T. 2006. Generalized empirical likelihood inference for nonlinear and time series models under weak identification. *Econ. Theory* 22:513–27
Otsu T. 2007. Penalized empirical likelihood estimation of semiparametric models. *J. Multivar. Anal.* 98:1923–54
Otsu T. 2011. Empirical likelihood estimation of conditional moment restriction models with unknown functions. *Econ. Theory* 27:8–46
Owen A. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75:237–49

Owen A. 1990. Empirical likelihood ratio confidence regions. *Ann. Stat.* 18:90–120
Owen A. 2001. *Empirical Likelihood*. New York: Chapman & Hall
Powell J, Stock JH, Stoker T. 1989. Semiparametric estimation of index coefficients. *Econometrica* 57:1403–30
Powell MJD. 1981. *Approximation Theory and Methods*. Cambridge, UK: Cambridge Univ. Press
Qin J, Lawless J. 1994. Empirical likelihood and general estimating equations. *Ann. Stat.* 22:300–25
Ramalho JJS. 2005. Small sample bias of alternative estimation methods for moment condition models: Monte Carlo evidence for covariance structures. *Stud. Nonlinear Dyn. Econom.* 9:1–20
Reiersøl O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9:l–24
Reiersøl O. 1945. Confluence analysis by means of instrumental sets of variables. *Ark. Mat. Astron. Fys.* 32A:1–119
Robinson PM. 1988a. Root-*N*-consistent semiparametric regression. *Econometrica* 56:931–54
Robinson PM. 1988b. The stochastic difference between econometric estimators. *Econometrica* 56:531–48
Sargan JD. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26:393–415
Sargan JD. 1959. The estimation of relationships with autocorrelated residuals by the use of the instrumental variables. *J. R. Stat. Soc. B* 21:91–105
Schennach SM. 2007. Point estimation with exponentially tilted empirical likelihood. *Ann. Stat.* 35:634–72
Shen X. 1997. On methods of sieves and penalization. *Ann. Stat.* 25:2555–91
Smith RJ. 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Econ. J.* 107:503–19
Smith RJ. 2000. Empirical likelihood estimation and inference. In *Applications of Differential Geometry to Econometrics*, ed. P Marriott, M Salmon, pp. 119–50. Cambridge, UK: Cambridge Univ. Press
Smith RJ. 2005. Automatic positive semi-definite HAC covariance matrix and GMM estimation. *Econ. Theory* 21:158–70
Smith RJ. 2007a. Efficient information theoretic inference for conditional moment restrictions. *J. Econom.* 138:430–60
Smith RJ. 2007b. Local GEL estimation with conditional moment restrictions. In *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*, ed. GDA Phillips, E Tzavalis, pp. 100–22. Cambridge, UK: Cambridge Univ. Press
Smith RJ. 2007c. Weak instruments and empirical likelihood: a discussion of the papers by D.W.K. Andrews and J.H. Stock and Y. Kitamura. In *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. 3, ed. RW Blundell, WK Newey, T Persson, pp. 238–60. Cambridge, UK: Cambridge Univ. Press
Smith RJ. 2011. GEL criteria for moment condition models. *Econ. Theory* 27:1192–235
Tripathi G, Kitamura Y. 2003. Testing conditional moment restrictions. *Ann. Stat.* 31:2059–95
Wright PG. 1928. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan
Wright S. 1925. *Corn and hog correlations*. Bull. 1300, US Dep. Agric., Washington, DC
Zhang J, Gijbels I. 2003. Sieve empirical likelihood and extensions of generalized least squares. *Scand. J. Stat.* 30:1–24