New Developments in Econometrics **Cemmap, UCL, June 2009**

Lecture 17, Thursday June 18th , 15.15-16.15

Generalized Method of Moments and Empirical Likelihood

1. Introduction

Generalized Method of Moments (henceforth GMM) estimation has become an important unifying framework for inference in econometrics in the last twenty years. It can be thought of as nesting almost all the common estimation methods such as maximum likelihood, ordinary least squares, instrumental variables and two–stage–least–squares and nowadays it is an important part of all advanced econometrics text books (Gallant, 1987; Davidson and McKinnon, 1993; Hamilton, 1994; Hayashi, 2000; Mittelhammer, Judge, and Miller, 2000; Ruud, 2000; Wooldridge, 2002). Its formalization by Hansen (1982) centers on the presence of known functions, labelled "moment functions", of observable random variables and unknown parameters that have expectation zero when evaluated at the true parameter values. The method generalizes the "standard" method of moments where expectations of known functions of observable random variables are equal to known functions of the unknown parameters. The "standard" method of moments can thus be thought of as a special case of the general method with the unknown parameters and observed random variables entering additively separable. The GMM approach links nicely to economic theory where orthogonality conditions that can serve as such moment functions often arise from optimizing behavior of agents. For example, if agents make rational predictions with squared error loss, their prediction errors should be orthogonal to elements of the information set. In the GMM framework the unknown parameters are estimated by setting the sample averages of these moment functions, the "estimating equations," as close to zero as possible.

The framework is sufficiently general to deal with the case where the number of moment functions is equal to the number of unknown parameters, the so–called "just–identified case", as well as the case where the number of moments exceeding the number of parameters to be estimated, the "over–identified case." The latter has special importance in economics where

the moment functions often come from the orthogonality of potentially many elements of the information set and prediction errors. In the just-identified case it is typically possible to estimate the parameter by setting the sample average of the moments exactly equal to zero. In the over-identified case this is not feasible. The solution proposed by Hansen (1982) for this case, following similar approaches in linear models such as two– and three–stage–least–squares, is to set a linear combination of the sample average of the moment functions equal to zero, with the dimension of the linear combination equal to the number of unknown parameters. The optimal linear combination of the moments depends on the unknown parameters, and Hansen suggested to employ initial, possibly inefficient, estimates to estimate this optimal linear combination. Chamberlain (1987) showed that this class of estimators achieves the semiparametric efficient bound given the set of moment restrictions. The Chamberlain paper is not only important for its substantive efficiency result, but also as a precursor to the subsequent empirical likelihood literature by the methods employed: Chamberlain uses a discrete approximation to the joint distribution of all the variables to show that the information matrix based variance bound for the discrete parametrization is equal to the variance of the GMM estimator if the discrete approximation is fine enough.

Th empirical likelihood literature developed partly in response to criticisms regarding the small sample properties of the two-step GMM estimator. Researchers found in a number of studies that with the degree of over-identification high, these estimators had substantial biases, and confidence intervals had poor coverage rates. See among others, Altonji and Segal (1996), Burnside and Eichenbaum (1996) , and Pagan and Robertson (1997). These findings are related to the results in the instrumental variables literature that with many or weak instruments two-stage-least squares can perform very badly (e.g., Bekker, 1994; Bound, Jaeger, and Baker, 1995; Staiger and Stock, 1997). Simulations, as well as theoretical results, suggest that the new estimators have LIML-like properties and lead to improved large sample properties, at the expense of some computational cost.

2. Examples

First the generic form of the GMM estimation problem in a cross–section context is

presented. The parameter vector θ^* is a K dimensional vector, an element of Θ, which is a subset of $\mathbb{R}^K$. The random vector Z has dimension P, with its support $\mathcal{Z}$ a subset of $\mathbb{R}^P$. The moment function, $\psi : \mathcal{Z} \times \Theta \to \mathbb{R}^M$, is a known vector valued function such that $E\left[\psi(Z, \theta^*)\right] = 0$, and $E\left[\psi(Z, \theta)\right] \neq 0$ for all $\theta \in \Theta$ with $\theta \neq \theta^*$. The researcher has available an independent and identically distributed random sample $Z_1, Z_2, \ldots, Z_N$. We are interested in the properties of estimators for θ^* in large samples.

Many, if not most models considered in econometrics fit this framework. Below are some examples, but this list is by no means exhaustive.

I. Maximum Likelihood

If one specifies the conditional distribution of a variable Y given another variable X as $f_{Y|X}(y|x, \theta)$, the score function satisfies these conditions for the moment function:

$$\psi(Y, X, \theta) = \frac{\partial \ln f}{\partial \theta}(Y|X, \theta).$$

By standard likelihood theory the score function has expectation zero only at the true value of the parameter. Interpreting maximum likelihood estimators as generalized method of moments estimators suggests a way of deriving the covariance matrix under misspecification (e.g., White, 1982), as well as an interpretation of the estimand in that case.

II. Linear Instrumental Variables

Suppose one has a linear model

$$Y = X'\theta^* + \varepsilon,$$

with a vector of instruments Z. In that case the moment function is

$$\psi(Y, X, Z, \theta) = Z' \cdot (Y - X'\theta).$$

The validity of Z as an instrument, together with a rank condition implies that θ^* is the unique solution to $E[\psi(Y, X, Z, \theta)] = 0$. This is a case where the fact that the methods allow for more moments than unknown parameters is of great importance as often instruments are independent of structural error terms, implying that any function of the basic instruments is orthogonal to the errors.

III. A Dynamic Panel Data Model

Consider the following panel data model with fixed effects:

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel data models studied in Keane and Runkle (1992), Chamberlain (1992), and Blundell and Bond (1998). This specific model has previously been studied by Bond, Bowsher, and Windmeijer (2001). One can construct moment functions by differencing and using lags as instruments, as in Arellano and Bond (1991), and Ahn and Schmidt, (1995):

$$\psi_{1t}(Y_{i1}, \ldots, Y_{iT}, \theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3, \ldots, T$, leading to a total of $(T-1) \cdot (T-2)/2$ moments, with only a single parameter. One would typically expect that the long lags do not necessarily contain much information, but they are often used to improve efficiency. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1}, \ldots, Y_{iT}, \theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

Despite the different nature of the two sets of moment functions, which makes them potentially very useful in the case that the autoregressive parameter is close to unity, they can all be combined in the GMM framework.

3. Two-step GMM Estimation

3.1 Estimation and Inference

In the just-identified case where M, the dimension of ψ, and K, the dimension of θ are identical, one can generally estimate θ^* by solving

$$0 = \frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \hat{\theta}_{\text{gmm}}). \tag{1}$$

If the sample average is replaced by the expectation, the unique solution is equal to θ^*, and under regularity conditions (e.g., Hansen, 1982, Newey and McFadden, 1994), solutions to (1) will be unique in large samples and consistent for θ^*. If $M > K$ the situation is more complicated as in general there will be no solution to (1).

Hansen's (1982) solution was to generalize the optimization problem to the minimization of the quadratic form

$$Q_{C,N}(\theta) = \frac{1}{N}\left[\sum_{i=1}^{N}\psi(z_i,\theta)\right]' \cdot C \cdot \left[\sum_{i=1}^{N}\psi(z_i,\theta)\right], \tag{2}$$

for some positive definite $M \times M$ symmetric matrix C. Under the regularity conditions given in Hansen (1982) and Newey and McFadden (1994), the minimand $\hat{\theta}_{\rm gmm}$ of (2) has the following large sample properties:

$$\hat{\theta}_{\rm gmm} \xrightarrow{p} \theta^*,$$

$$\sqrt{N}(\hat{\theta}_{\rm gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C \Gamma)^{-1}\Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}),$$

where

$$\Delta = \mathbb{E}\left[\psi(Z_i,\theta^*)\psi(Z_i,\theta^*)'\right] \qquad \text{and} \quad \Gamma = \mathbb{E}\left[\frac{\partial}{\partial\theta'}\psi(Z_i,\theta^*)\right].$$

In the just–identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial, as $\hat{\theta}_{\rm gmm}$ will, at least in large samples, be equal to the value of θ that sets the average moments exactly equal to zero. In that case Γ is a square matrix, and because it is full rank by assumption, Γ is invertible and the asymptotic covariance matrix reduces to $(\Gamma'\Delta^{-1}\Gamma)^{-1}$, irrespective of the choice of C. In the overidentified case with $M > K$, however, the choice of the weight matrix C is important. The optimal choice for C in terms of minimizing the asymptotic variance is in this case the inverse of the covariance of the moments, Δ^{-1}. Using the optimal weight matrix, the asymptotic distribution is

$$\sqrt{N}(\hat{\theta}_{\rm gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1}). \tag{3}$$

This estimator is generally not feasible because typically Δ^{-1} is not known to the researcher. The feasible solution proposed by Hansen (1982) is to obtain an initial consistent, but generally inefficient, estimate of θ^* by minimizing $Q_{C,N}(\theta)$ using an arbitrary positive definite $M \times M$ matrix C, e.g., the identity matrix of dimension M. Given this initial estimate, $\tilde{\theta}$, one can estimate the optimal weight matrix as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N} \sum_{i=1}^{N} \psi(z_i, \tilde{\theta}) \cdot \psi(z_i, \tilde{\theta})' \right]^{-1}.$$

In the second step one estimates θ^* by minimizing $Q_{\hat{\Delta}^{-1},N}(\theta)$. The resulting estimator $\hat{\theta}_{\rm gmm}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Hansen (1982) also suggested a specification test for this model. If the number of moments exceeds the number of free parameters, not all average moments can be set equal to zero, and their deviation from zero forms the basis of Hansen's test, similar to tests developed by Sargan (1958). See also Newey (1985a, 1985b). Formally, the test statistic is

$$T = Q_{\hat{\Delta},N}(\hat{\theta}_{\rm gmm}).$$

Under the null hypothesis that all moments have expectation equal to zero at the true value of the parameter, θ^*, the distribution of the test statistic converges to a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions, $M - K$.

One can also interpret the two–step estimator for over–identified GMM models as a just–identified GMM estimator with an augmented parameter vector (e.g., Newey and McFadden, 1994; Chamberlain and Imbens, 1995). Define the following moment function:

$$h(x,\delta) = h(x,\theta,\Gamma,\Delta,\beta,\Lambda) = \begin{pmatrix} \Lambda - \frac{\partial \psi}{\partial \theta'}(x,\beta) \\ \Lambda' C \psi(x,\beta) \\ \Delta - \psi(x,\beta)\psi(x,\beta)' \\ \Gamma - \frac{\partial \psi}{\partial \theta'}(x,\theta) \\ \Gamma' \Delta^{-1} \psi(x,\theta) \end{pmatrix}. \tag{4}$$

Because the dimension of the moment function $h(\cdot)$, $M\times K+K+(M+1)\times M/2+M\times K+K = (M+1)\times(2K+M/2)$, is equal to the combined dimensions of its parameter arguments, the

estimator for $\delta = (\theta, \Gamma, \Delta, \beta, \Lambda)$ obtained by setting the sample average of $h(\cdot)$ equal to zero is a just–identified GMM estimator. The first two components of $h(x, \delta)$ depend only on β and Λ, and have the same dimension as these parameters. Hence β^* and Λ^* are implicitly defined by the equations

$$E\left[\begin{pmatrix} \Lambda - \frac{\partial \psi}{\partial \theta'}(X,\beta) \\ \Lambda' C \psi(X,\beta) \end{pmatrix}\right] = 0.$$

Given β^* and Λ^*, Δ^* is defined through the third component of $h(x, \delta)$, and given β^*, Λ^* and Δ^* the final parameters θ^* and Γ^* are defined through the last two moment functions.

This interpretation of the over-identified two-step GMM estimator as a just-identified GMM estimator in an augmented model is interesting because it also emphasizes that results for just–identified GMM estimators such as the validity of the bootstrap can directly be translated into results for over–identified GMM estimators. In another example, using the standard approach to finding the large sample covariance matrix for just–identified GMM estimators one can use the just-identified representation to find the covariance matrix for the over–identified GMM estimator that is robust against misspecification: the appropriate submatrix of

$$\left(E\left[\frac{\partial h}{\partial \delta}(X,\delta^*)\right]\right)^{-1} E[h(Z,\delta^*)h(Z,\delta^*)'] \left(E\left[\frac{\partial h}{\partial \delta}(Z,\delta^*)\right]\right)^{-1},$$

estimated by averaging at the estimated values. This is the GMM analogue of the White (1982) covariance matrix for the maximum likelihood estimator under misspecification.

3.2 Efficiency

Chamberlain (1987) demonstrated that Hansen's (1982) estimator is efficient, not just in the class of estimators based on minimizing the quadratic form $Q_{N,C}(\theta)$, but in the larger class of semiparametric estimators exploiting the full set of moment conditions. What is particularly interesting about this argument is the relation to the subsequent empirical likelihood literature. Many semiparametric efficiency bound arguments (e.g., Newey, 1991; Hahn, 1994) implicitly build fully parametric models that include the semiparametric one and then search for the least favorable parametrization. Chamberlain's argument is qualitatively different.

He proposes a specific parametric model that can be made arbitrarily flexible, and thus arbitrarily close to the model that generated the data, but does not typically include that model. The advantage of the model Chamberlain proposes is that it is in some cases very convenient to work with in the sense that its variance bound can be calculated in a straightforward manner. The specific model assumes that the data are discrete with finite support $\{\lambda_1, \ldots, \lambda_L\}$, and unknown probabilities $\pi_1, \ldots, \pi_L$. The parameters of interest are then implicitly defined as functions of these points of support and probabilities. With only the probabilities unknown, the variance bound on the parameters of the approximating model are conceptually straightforward to calculate. It then sufficies to translate that into a variance bound on the parameters of interest. If the original model is over-identified, one has restrictions on the probabilities. These are again easy to evaluate in terms of their effect on the variance bound.

Given the discrete model it is straightforward to obtain the variance bound for the probabilities, and thus for any function of them. The remarkable point is that one can rewrite these bounds in a way that does not involve the support points. This variance turns out to be identical to the variance of the two-step GMM estimator, thus proving its efficiency.

4. Empirical Likelihood

4.1 Background

To focus ideas, consider a random sample $Z_1, Z_2, \ldots, Z_N$, of size N from some unknown distribution. If we wish to estimate the common distribution of these random variables, the natural choice is the empirical distribution, that puts weight $1/N$ on each of the N sample points. However, in a GMM setting this is not necessarily an appropriate estimate. Suppose the moment function is

$$\psi(z, \theta) = z,$$

implying that the expected value of Z is zero. Note that in this simple example this moment function does not depend on any unknown parameter. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$ as $E_{\hat{F}_{emp}}[Z] = \sum z_i/N \neq 0$.

The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\hat{F}$ does satisfy the restriction. In other words, the approach is to look for the distribution function closest to $\hat{F}_{\rm emp}$, within the set of distribution functions satisfying $E_F[Z] = 0$. Empirical likelihood provides an operationalization of the concept of closeness here. The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i,$$

for $0 \leq \pi_i \leq 1$, $\sum_{i=1}^{N} \pi_i = 1$. This is not a likelihood function in the standard sense, and thus does not have all the properties of likelihood functions. The empirical likelihood estimator for the distribution function is

$$\max_{\pi} \sum_{i=1}^{N} \pi_i \qquad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i z_i = 0.$$

Without the second restriction the π's would be estimated to be $1/N$, but the second restriction forces them slightly away from $1/N$ in a way that ensures the restriction is satisfied. In this example the solution for the Lagrange multiplier is the solution to the equation

$$\sum_{i=1}^{N} \frac{z_i}{1 + t \cdot z_i} = 0,$$

and the solution for π_i is:

$$\hat{\pi}_i = 1/(1 + t \cdot z_i).$$

More generally, in the over-identified case a major focus is on obtaining point estimates through the following estimator for θ:

$$\max_{\theta, \pi} \sum_{i=1}^{N} \ln \pi, \qquad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0. \tag{5}$$

Qin and Lawless (1994) and Imbens (1997) show that this estimator is equivalent, to order $O_p(N^{-1/2})$, to the two-step GMM estimator. This simple discussion illustrates that for some, and in fact many, purposes the empirical likelihood has the same properties as a parametric likelihood function. This idea, first proposed by Owen (1988), turns out to be very powerful

with many applications. Owen (1988) shows how one can construct confidence intervals and hypothesis tests based on this notion.

Related ideas have shown up in a number of places. Cosslett's (1981) work on choice-based sampling can be interpreted as maximizing a likelihood function that is the product of a parametric part coming from the specification of the conditional choice probabilities, and an empirical likelihood function coming from the distribution of the covariates. See Imbens (1992) for a connection between Cosslett's work and two-step GMM estimation. As mentioned before, Chamberlain's (1987) efficiency proof essentially consists of calculating the distribution of the empirical likelihood estimator and showing its equivalence to the distribution of the two-step GMM estimator. See Back and Brown (1990) and Kitamura and Stutzer (1997) for a discussion of the dependent case.

4.2 Cressie-Read Discrepancy Statistics and Generalized Empirical Likelihood

In this section we consider a generalization of the empirical likelihood estimators based on modifications of the objective function. Corcoran (1998) (see also Imbens, Spady and Johnson, 1998), focus on the Cressie-Read discrepancy statistic, for fixed λ, as a function of two vectors p and q of dimension N (Cressie and Read 1984):

$$I_\lambda(p,q) = \frac{1}{\lambda\cdot(1+\lambda)}\sum_{i=1}^N p_i\left[\left(\frac{p_i}{q_i}\right)^\lambda - 1\right].$$

The Cressie-Read minimum discrepancy estimators are based on minimizing this difference between the empirical distribution, that is, the N-dimensional vector with all elements equal to $1/N$, and the estimated probabilities, subject to all the restrictions being satisfied.

$$\min_{\pi,\theta} I_\lambda(\iota/N,\pi) \qquad \text{subject to } \sum_{i=1}^N \pi_i = 1, \quad \text{and } \sum_{i=1}^N \pi_i\cdot\psi(z_i,\theta) = 0.$$

If there are no binding restrictions, because the dimension of $\psi(\cdot)$ and θ agree (the just-identified case), the solution for π is the empirical distribution it self, and $\pi_i = 1/N$. More generally, if there are over-identifying restrictions, there is no solution for θ to $\sum_i \psi(z_i,\theta)/N = 0$, and so the solution for π_i is as close as possible to $1/N$ in a way that ensures there is

an exact solution to $\sum_i \pi_i \psi(z_i,\theta) = 0$. The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Three special cases of this class have received most attention. First, the empirical likelihood estimator itself, which can be interpreted as the case with $\lambda \to 0$. This has the nice interpretation that it is the exact maximum likelihood estimator if Z has a discrete distribution. It does not rely on the discreteness for its general properties, but this interpretation does suggest that it may have attractive large sample properties.

The second case is the exponential tilting estimator with $\lambda \to -1$ (Imbens, Spady and Johnson, 1998), whose objective function is equal to the empirical likelihood objective funtion with the role of π and ι/N reversed. It can also be written as

$$\min_{\pi,\theta} \sum_{i=1}^N \pi_i \ln \pi_i \qquad \text{subject to } \sum_{i=1}^N \pi_i = 1, \quad \text{and } \sum_{i=1}^N \pi_i \psi(z_i,\theta) = 0.$$

Third, the case with $\lambda = -2$. This case was originally proposed by Hansen, Heaton and Yaron (1996) as the solution to

$$\min_\theta \frac{1}{N}\left[\sum_{i=1}^N \psi(z_i,\theta)\right]' \cdot \left[\frac{1}{N}\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\right]^{-1} \cdot \left[\sum_{i=1}^N \psi(z_i,\theta)\right],$$

where the GMM objective function is minimized over the θ in the weight matrix as well as the θ in the average moments. Hansen, Heaton and Yaron (1996) labeled this the continuously updating estimator. Newey and Smith (2004) pointed out that this estimator fits in the Cressie-Read class.

Smith (1997) considers a more general class of estimators, which he labels generalized empirical likelihood estimators, starting from a different perspective. For a given function $g(\cdot)$, normalized so that it satisfied $g(0) = 1$, $g'(0) = 1$, consider the saddle point problem

$$\max_\theta \min_t \sum_{i=1}^N g(t'\psi(z_i,\theta)).$$

This representation is more attractive from a computational perspective, as it reduces the dimension of the optimization problem to $M + K$ rather than a constrained optimization

problem of dimension $K+N$ with $M+1$ restrictions. There is a direct link between the t parameter in the GEL representation and the Lagrange multipliers in the Cressie-Read representation. Newey and Smith (2004) how to choose $g(\cdot)$ for a given λ so that the corresponding GEL and Cressie-Read estimators agree.

In general the differences between the estimators within this class is relatively small compared to the differences between them and the two-step GMM estimators. In practice the choice between them is largely driven by computational issues, which will be discussed in more detail in Section 5. The empirical likelihood estimator does have the advantage of its exact likelihood interpretation and the resulting optimality properties for its bias-corrected version (Newey and Smith, 2004). On the other hand, Imbens, Spady and Johnson (1998) argue in favor of the exponential tilting estimator as its influence function stays bounded where as denominator in the probabilities in the empirical likelihood estimator can get large. In simulations researcher have encountered more convergence problems with the continuously updating estimator (e.g., Hansen, Heaton and Yaron, 1996; Imbens, Johnson and Spady, 1998).

4.3 Testing

Associated with the empirical likelihood estimators are three tests for over-identiyfing restrictions, similar to the classical trinity of tests, the likelihood ratio, the Wald, and the Lagrange multiplier tests. Here we briefly review the implementation of the three tests in the empirical likelihood context. The leading terms of all three tests are identical to that of the test developed by Hansen (1982) based on the quadratic form in the average moments.

The first test is based on the value of the empirical likelihood function. The test statistic compares the value of the empirical likelihood function at the restricted estimates, the $\hat{\pi}_i$ with that at the unrestricted values, $\pi_i = 1/N$:

$$LR = 2 \cdot (L(\iota/N) - L(\hat{\pi})), \qquad \text{where} \quad L(\pi) = \sum_{i=1}^{N} \ln \pi_i.$$

As in the parametric case, the difference between the restricted and unrestricted likelihood function is multiplied by two to obtain, under regularity conditions, e.g., Newey and Smith

(2004), a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions for the test statistic under the null hypothesis.

The second test, similar to Wald tests, is based on the difference between the average moments and their probability limit under the null hypothesis, zero. As in the standard GMM test for overidentifying restrictions (Hansen, 1982), the average moments are weighted by the inverse of their covariance matrix:

$$Wald = \frac{1}{N}\left[\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\right]' \hat{\Delta}^{-1}\left[\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\right],$$

where $\hat{\Delta}$ is an estimate of the covariance matrix

$$\Delta = E[\psi(Z,\theta^*)\psi(Z,\theta)'],$$

typically based on a sample average at some consistent estimator for θ^*:

$$\hat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\psi(z_i,\hat{\theta})',$$

or sometimes a fully efficient estimator for the covariance matrix,

$$\hat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_i\psi(z_i,\hat{\theta})\psi(z_i,\hat{\theta})',$$

The standard GMM test uses an initial estimate of θ^* in the estimation of Δ, but with the empirical likelihood estimators it is more natural to substitute the empirical likelihood estimator itself. The precise properties of the estimator for Δ do not affect the large sample properties of the test, and like the likelihood ratio test, the test statistic has in large samples a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions.

The third test is based on the Lagrange multipliers t. In large samples their variance is

$$V_t = \Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}.$$

This matrix is singular, with rank equal to $M-K$. One option is therefore to compare the Lagrange multipliers to zero using a generalized inverse of their covariance matrix:

$$LM_1 = t'\left(\Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\right)^{-g} t.$$

This is not very attractive, as it requires the choice of a generalized inverse. An alternative is to use the inverse of Δ^{-1} itself, leading to the test statistic

$$LM_2 = t'\Delta t.$$

Because

$$\sqrt{N}\cdot t = V_t \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta^*) + o_p(1),$$

and $V_t\Delta V_t = V_t V_t^{-g} V_t = V_t$, it follows that

$$LM_2 = LM_1 + o_p(1).$$

Imbens, Johnson and Spady (1998) find in their simulations that tests based on LM_2 perform better than those based on LM_1. In large samples both have a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions. Again we can use this test with any efficient estimator for t, and with the Lagrange multipliers based on any of the discrepancy measures.

Imbens, Spady and Johnson (1998), and Bond, Bowsher and Windmeijer (2001) investigate through simulations the small sample properties of various of these tests. It appears that the Lagrange multiplier tests are often more attractive than the tests based on the average moments, although there is so far only limited evidence in specific models. One can use the same ideas for constructing confidence intervals that do not directly use the normal approximation to the sampling distribution of the estimator. See for discussions Smith (1998) and Imbens and Spady (2002).

6. Computational Issues

The two-step GMM estimator requires two minimizations over a K-dimensional space. The empirical likelihood estimator in its original likelihood form (5) requires maximization over a space of dimension K (for the parameter θ) plus N (for the N probabilities), subject to $M+1$ restrictions (on the M moments and the adding up restriction for the probabilities). This is in general a much more formidable computational problem than two optimizations

in a K-dimensional space. A number of approaches have been attempted to simplify this problem. Here we disuss three of them in the context of the exponential tilting estimator, although most of them directly carry over to other members of the Cressie-Read or GEL classes.

6.1 Solving the First Order Conditions

The first approach we discuss is focuses on the first order conditions and then concentrates out the probabilities π. This reduces the problem to one of dimension $K+M$, K for the parameters of interest and M for the Lagrange multipliers for the restrictions, which is is clearly a huge improvement, as the dimension of the problem no longer increases with the sample size. Let μ and t be the Lagrange multipliers for the restrictions $\sum \pi_i = 1$ and $\sum \pi_i \psi(z_i,\theta) = 0$. The first order conditions for the π's and θ and the Lagrange multipliers are

$$
\begin{aligned}
0 &= \ln \pi_i - 1 - \mu + t'\psi(z_i,\theta), \\
0 &= \sum_{i=1}^N \pi_i \frac{\partial \psi}{\partial \theta'}(z_i,\theta), \\
0 &= \exp(\mu - 1) \sum_{i=1}^N \exp\left(t'\psi(z_i,\theta)\right), \\
0 &= \exp(\mu - 1) \sum_{i=1}^N \psi(z_i,\theta) \cdot \exp\left(t'\psi(z_i,\theta)\right).
\end{aligned}
$$

The solution for π is

$$
\pi_i = \exp(\mu - 1 + t'\psi(z_i,\theta)).
$$

To determine the Lagrange multipliers t and the parameter of interest θ we only need π_i up to a constant of proportionality, so we can solve

$$
0 = \sum_{i=1}^N \psi(z_i,\theta) \exp(t'\psi(z_i,\theta)), \tag{6}
$$

and

$$
0 = \sum_{i=1}^N t' \frac{\partial \psi}{\partial \theta}(z_i,\theta) \exp(t'\psi(z_i,\theta)) \tag{7}
$$

Solving the system of equations (6) and (7) is not straightforward. Because the probability limit of the solution for t is zero, the derivative with respect to θ of both first order conditions converges zero. Hence the matrix of derivatives of the first order conditions converges to a singular matrix. As a result standard approaches to solving systems of equations can behave erratically, and this approach to calculating $\hat{\theta}$ has been found to have poor operating characteristics.

6.2 Penalty Function Approaches

Imbens, Spady and Johnson (1998) characterize the solution for θ and t as

$$\max_{\theta,t} K(t,\theta) \qquad \text{subject to } K_t(t,\theta)=0, \tag{8}$$

where $K(t,\theta)$ is the empirical analogue of the cumulant generating function:

$$K(t,\theta)=\ln\left[\frac{1}{N}\sum_{i=1}^{N}\exp(t'\psi(z_i,\theta)\right].$$

They suggest solving this optimization problem by maximizing the unconstrained objective function with a penalty term that consists of a quadratic form in the restriction:

$$\max_{\theta,t} K(t,\theta)-0.5\cdot A\cdot K_t(t,\theta)'W^{-1}K_t(t,\theta), \tag{9}$$

for some positive definite $M\times M$ matrix W, and a positive constant A. The first order conditions for this problem are

$$0=K_\theta(t,\theta)-A\cdot K_{t\theta}(t,\theta)W^{-1}K_t(t,\theta),$$

$$0=K_t(t,\theta)-A\cdot K_{tt}(t,\theta)W^{-1}K_t(t,\theta).$$

For A large enough the solution to this unconstrained maximization problem is identical to the solution to the constrained maximization problem (8). This follows from the fact that the constraint is in fact the first order condition for $K(t,\theta)$. Thus, in contrast to many penalty function approaches, one does not have to let the penalty term go to infinity to obtain the solution to the constrained optimization problem, one only needs to let the penalty term

increase sufficiently to make the problem locally convex. Imbens, Spady and Johnson (1998) suggest choosing

$$W = K_{tt}(t,\theta) + K_t(t,\theta)K_t(t,\theta)',$$

for some initial values for t and θ as the weight matrix, and report that estimates are generally not sensitive to the choices of t and θ.

6.3 Concentrating out the Lagrange Multipliers

Mittelhammer, Judge and Schoenberg (2001) suggest concentrating out both probabilities and Lagrange multipliers and then maximizing over θ without any constraints. As shown above, concentrating out the probabilities π_i can be done analytically. Although it is not in general possible to solve for the Lagrange multipliers t analytically, other than in the continuously updating case, for given θ it is easy to numerically solve for t. This involves solving, in the exponential tilting case,

$$\min_t \sum_{i=1}^N \exp(t'\psi(z_i,\theta)).$$

This function is strictly convex as a function of t, with the easy to calculate first and second derivatives equal to

$$\sum_{i=1}^N \psi(z_i,\theta)\exp(t'\psi(z_i,\theta)),$$

and

$$\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\exp(t'\psi(z_i,\theta)),$$

respectively. Therefore concentrating out the Lagrange multipliers is computationally fast using a Newton-Raphson algorithm. The resulting function $t(\theta)$ has derivatives with respect to θ equal to:

$$\frac{\partial t}{\partial \theta'}(\theta) = -\left(\frac{1}{N}\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\exp(t(\theta)'\psi(z_i,\theta)\right)^{-1}$$

$$\cdot \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \psi}{\partial \theta'}(z_i, \theta) \exp(t(\theta)'\psi(z_i, \theta) + \psi(z_i, \theta) t(\theta)' \frac{\partial \psi}{\partial \theta'}(z_i, \theta) \exp(t(\theta)'\psi(z_i, \theta)) \right)$$

After solving for $t(\theta)$, one can solve

$$\max_{\theta} \sum_{i=1}^{N} \exp(t(\theta)'\psi(z_i, \theta)). \tag{10}$$

Mittelhammer, Judge, and Schoenberg (2001) use methods that do not require first derivatives to solve (10). This is not essential. Calculating first derivatives of the concentrated objective function only requires first derivatives of the moment functions, both directly and indirectly through the derivatives of $t(\theta)$ with respect to θ. In general these are straightforward to calculate and likely to improve the performance of the algorithm.

In this method in the end the researcher only has to solve one optimization in a K-dimensional space, with the provision that for each evaluation of the objective function one needs to numerically evaluate the function $t(\theta)$ by solving a convex maximization problem. The latter is fast, especially in the exponential tilting case, so that although the resulting optimization problem is arguably still more difficult than the standard two-step GMM problem, in practice it is not much slower. In the simulations below I use this method for calculating the estimates. After concentrating out the Lagrange multipliers using a Newton-Rahpson algorithm that uses both first and second derivatives, I use a Davidon-Fletcher-Powell algorithm to maximize over θ, using analytic first derivatives. Given a direction I used a line search algoritm based on repeated quadratic approximations.

7. A Dynamic Panel Data Model

To get a sense of the finite sample properties of the empirical likelihood estimators we compare some of the GMM methods in the context of the panel data model briefly discussed in Section 2, using some simulation results from Imbens. The model is

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel

data models extensively studied in the literature. Bond, Bowsher and Windmeijer (2001) study this and similar models to evaluate the performance of test statistics based on different GMM and gel estimators. We use the moments

$$\psi_{1t}(Y_{i1},\ldots,Y_{iT},\theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3,\ldots,T$, leading to a total of $(T-1)\cdot(T-2)/2$ moments. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1},\ldots,Y_{iT},\theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

It is important to note, given the results discussed in Section 4, that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. As a result there is potentially a substantial difference between the different estimators, especially when the degree of overidentification is high.

We report some simulations for a data generating process with parameter values estimated on data from Abowd and Card (1989) taken from the PSID. See also Card (1994). This data set contains earnings data for 1434 individuals for 11 years. The individuals are selected on having positive earnings in each of the eleven years, and we model their earnings in logarithms. We focus on estimation of the autoregressive coefficient θ.

We then generate artificial data sets to investigate the repeated sampling properties of these estimators. Two questions are of most interest. First, how do the median bias and median-absolute-error deteriorate as a function of the degree of over-identification? Here, unlike in the theoretical discussion in Section 4, the additional moments, as we increase the number of years in the panel, do contain information, so they may in fact increase precision, but at the same time one would expect based on the theoretical calculations that the accuracy of the asymptotic approximations for a fixed sample size deteriorates with the

number of years. Second, we are interested in the performance of the confidence intervals for the parameter of interest. In two-stage-least-squares settings it was found that with many weak instruments the performance of standard confidence intervals varied widely between liml and two-stage-least-squares estimators. Given the analogy drawn by Hansen, Heaton and Yaron (1996) between the continuously updating estimator and liml, the question arises how the confidence intervals differ between two-step GMM and the various Cressie-Read and GEL estimators.

Using the Abowd-Card data we estimate θ and the variance of the fixed effect and the idiosyncratic error term. The latter two are estimated to be around 0.3. We then consider data generating processes where the individual effect η_i has mean zero and standard deviation equal to 0.3, and the error term has mean zero and standard deviation 0.3. We $\theta = 0.9$ in the simulations. This is larger than the value in estimated from the Abowd-Card data. We compare the standard Two-Step GMM estimator and the Exponential Tilting Estimator. Table 1 contains the results. With the high autoregressive coefficient, $\theta = 0.9$, the two-step GMM estimator has substantial bias and poor coverage rates.The exponential tilting estimator does much better with the high autoregressive coefficient. The bias is small, on the order of 10% of the standard error, and the coverage rate is much closer to the nominal one.

References

Abowd, J. and D. Card, (1989), "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57 (2), 441-445.

Ahn, S., and P. Schmidt, (1995), "Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, 68, 5-28.

Altonji, J., and L. Segal, (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, Vol 14, No. 3, 353–366.

Back, K., and D. Brown, (1990), "Estimating Distributions from Moment Restrictions", working paper, Graduate School of Business, Indiana University.

Bekker, P., (1994), "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 62, 657-681.

Bond, S., C. Bowsher, and F. Windmeijer, (2001), "Criterion-based Inference for GMM in Linear Dynamic Panel Data Models", IFS, London.

Bound, J., D. Jaeger, and R. Baker, (1995), "Problems with Instrumental Variables Estimation when the Correlation between Instruments and the Endogenous Explanatory Variable is Weak", forthcoming, em Journal of the American Statistical Association.

Burnside, C., and M., Eichenbaum, (1996), "Small Sample Properties of Generalized Method of Moments Based Wald Tests", *Journal of Business and Economic Statistics*, Vol. 14, 294-308.

Card, D., (1994) "Intertemporal Labour Supply: an Assessment", in: *Advances in Econometrics*, Simms (ed), Cambridge University Press.

Chamberlain, G., (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, vol. 34, 305–334, 1987

Corcoran, S., (1998), "Bartlett Adjustment of Empirical Discrepancy Statistics", *Biometrika*.

Cosslett, S. R., (1981), "Maximum Likelihood Estimation for Choice–based Samples", *Econometrica*, vol. 49, 1289–1316,

Cressie, N., and T. Read, (1984), "Multinomial Goodness-of-Fit Tests", *Journal of the Royal Statistical Society*, Series B, 46, 440-464.

Hall, A., (2005), *Generalized Method of Moments*, Oxford University Press.

Hansen, L–P., (1982), "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, vol. 50, 1029–1054.

Hansen, L.–P., J. Heaton, and A. Yaron, (1996), "Finite Sample Properties of Some Alternative GMM Estimators", *Journal of Business and Economic Statistics*, Vol 14, No. 3, 262–280.

Imbens, G. W., (1992), "Generalized Method of Moments and Empirical Likelihood," *Journal of Business and Economic Statistics*, vol. 60.

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

Imbens, G., and R. Spady, (2002), "Confidence Intervals in Generalized Method of Moments Models," *Journal of Econometrics*, 107, 87-98.

Imbens, G., and R. Spady, (2005), "The Performance of Empirical Likelihood and Its Generalizations," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Andrews and Stock (eds).

Kitamura, Y., and M. Stutzer, (1997), "An Information-theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica*, Vol. 65, 861-874.

Mittelhammer, R., G. Judge, and R. Schoenberg, (2005), "Empirical Evidence Concerning the Finite Sample Performance of EL-Type Structural Equation Estimation and Inference Methods," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Andrews and Stock (eds).

Mittelhammer, R., G. Judge, and D. Miller, (2000), *Econometric Foundations*, Cambridge University Press, Cambridge.

Newey, W., (1985), "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, vol. 29, 229–56.

Newey, W., and D. McFadden, (1994) "Estimation in Large Samples", in: McFadden and Engle (Eds.), *The Handbook of Econometrics*, Vol. 4.

Newey, W., and R. Smith, (2004), "Higher Order Properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 573-595.

Owen, A., (1988), "Empirical Likelihood Ratios Confidence Intervals for a Single Functional", *Biometrika*, 75, 237-249.

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

Pagan, A., and J. Robertson, (1997), "GMM and its Problems", unpublished manuscript, Australian National University.

Qin, and J. Lawless, (1994), "Generalized Estimating Equations", *Annals of Stat.*

Smith, R., (1997), "Alternative Semiparametric Likelihood Approaches to Generalized Method of Moments Estimation", *Economic Journal*, 107, 503-19.

Staiger, D., and J. Stock, (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557-586.

White, H., (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, vol. 50, 1–25.

Wooldridge, J., (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples", *Econometrica* 67, No. 6, 1385-1406.

Table 1: SIMULATIONS, $\theta = 0.9$

	Number of time periods								
	3	4	5	6	7	8	9	10	11
Two-Step GMM									
median bias	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
relative median bias	-0.02	0.08	0.03	0.08	0.03	0.11	0.08	0.13	0.11
median absolute error	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.88	0.85	0.82	0.80	0.80	0.79	0.78	0.79	0.76
covarage rate 95% ci	0.92	0.91	0.89	0.87	0.85	0.86	0.86	0.88	0.84
Exponential Tilting									
median bias	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00
relative median bias	0.04	0.09	0.02	-0.00	0.01	0.01	-0.02	0.08	0.13
median absolute error	0.05	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.87	0.86	0.84	0.86	0.88	0.86	0.87	0.88	0.87
covarage rate 95% ci	0.91	0.90	0.90	0.91	0.93	0.92	0.91	0.93	0.93

The relative median bias reports the bias divided by the large sample standard error. All results based on 10,000 replications.