"New Developments in Econometrics"

Lecture 17

Generalized Method of Moments

and Empirical Likelihood

Guido Imbens

Cemmap Lectures, UCL, June 2009

Outline

1. Introduction

2. Generalized Method of Moments Estimation

3. Empirical Likelihood

4. Computational Issues

5. A Dynamic Panel Data Model

1. Introduction

GMM has provided a very influential framework for estimation since Hansen (1982). Many models and estimators fit in.

In the case with over-identification the traditional approach is to use a two-step method with estimated weight matrix.

For this case Empirical Likelihood provides attractive alternative with higher order bias properties, and liml-like advantages in settings with high degrees of over-identification.

The choice between various EL-type estimators is less important than the choice between the class and two-step gmm.

Computationally the estimators are only marginally more demanding. Most effective seems to be to concentrate out Lagrange multipliers.

2. Generalized Method of Moments Estimation

Generic form of the GMM estimation problem: 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

$$\mathbb{E}\left[\psi(Z, \theta^*)\right] = 0, \quad \text{and} \quad \mathbb{E}\left[\psi(Z, \theta)\right] \neq 0, \text{ for all } \theta \neq \theta^*$$

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.

Example 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.

Example 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 $E[\psi(Y, X, Z, \theta)] = 0$.

Example 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 (θ).

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:

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

GMM: Estimation

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

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

Under regularity conditions solutions will be unique in large samples and consistent for θ^*. If $M > K$ there is in general there will be no solution to (1).

Hansen's solution was to minimize the quadratic form

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

for some positive definite $M \times M$ symmetric matrix C (which if $M = K$ still leads to a $\widehat{\theta}$ that solves the equation (1).

GMM: Large Sample Properties

Under regularity conditions the minimand $\widehat{\theta}_{\text{gmm}}$ has the following large sample properties:

$$\widehat{\theta}_{\text{gmm}} \xrightarrow{p} \theta^*,$$

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

where

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

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.

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.

GMM: Optimal Weight Matrix

In the overidentified case with $M > K$ 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}.

Then:

$$\sqrt{N}(\widehat{\theta}_{\text{gmm}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' \Delta^{-1} \Gamma)^{-1}). \tag{2}$$

This estimator is not feasible because Δ^{-1} is unknown.

The feasible solution is to obtain an initial consistent, but generally inefficient, estimate of θ^* and then 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 $\widehat{\theta}_{\text{gmm}}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Compare to TSLS having the same asymptotic distribution as estimator with optimal instrument.

GMM: Specification Testing

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 a test.Formally, the test statistic is

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

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

Interpreting Over-identified GMM as a Just-identified Moment Estimator

One can also interpret the two–step estimator for over–identified GMM models as a just–identified GMM estimator with an augmented parameter vector. Fix an arbitrary $M \times M$ postitive definite matrix C. Then:

$$h(x,\delta) = h(x,\theta,\Gamma,\Delta,\beta,\Lambda) = \begin{pmatrix} \Lambda - \frac{\partial \psi}{\partial \theta'}(x,\beta) \\ \Lambda' C \psi(x,\beta) \\ \Delta - \psi(x,\beta)\psi(x,\beta)' \\ \Gamma - \frac{\partial \psi}{\partial \theta'}(x,\theta) \\ \Gamma' \Delta^{-1} \psi(x,\theta) \end{pmatrix}. \quad (3)$$

This interpretation 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.

For example, 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.

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.

Chamberlain 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 Cramér-Rao variance bound is conceptually straightforward to calculate.

It turns out this is equal to variance of GMM estimator with optimal weight matrix.

3. Empirical Likelihood

Consider a random sample $Z_1, Z_2, \ldots, Z_N$, of size N from some unknown distribution. The natural choice for estimating the dsitribution function is the empirical distribution, that puts weight $1/N$ on each of the N sample points.

Suppose we also know that $\mathbb{E}[Z] = 0$. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$ as $E_{\widehat{F}_{emp}}[Z] = \sum z_i/N \neq 0$.

The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\widehat{F}$ does satisfy the restriction.

The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i, \quad \text{for } 0 \leq \pi_i \leq 1, \ \sum_{i=1}^{N} \pi_i = 1$$

The empirical likelihood estimator for the distribution function is, given $\mathbb{E}[Z] = 0$,

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

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.

This leads to

$$\widehat{\pi}_i = 1/(1 + t \cdot z_i) \quad \text{where t solves } \textstyle\sum_{i=1}^{N} \frac{z_i}{1+t \cdot z_i} = 0,$$

EL: The General Case

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

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

This is equivalent, to first order asymptotics, to the two-step GMM estimator.

For many purposes the empirical likelihood has the same properties as a parametric likelihood function. (Qin and Lawless, 1994; Imbens, 1997; Kitamura and Stutzer, 1997).

EL: Cressie-Read Discrepancy Statistics

Define

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

and solve

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

The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Empirical Likelihood is special case with $\lambda_{\mathsf{EL}} \longrightarrow 0$.

EL: Generalized Empirical Likelihood

Smith (1997), Newey and Smith (1994) considers a more general class of estimators. For a given function $g(\cdot)$, normalized so that it satisfied $g'(0) = -1$, $g''(0) = -1$, $g'''(o) = \rho$, consider the saddle point problem

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

This representation is 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: $\lambda = \rho + 2$, and $\rho_{\text{EL}} = -2$

GEL: Special cases, Continuously Updating Estimator

$\lambda = -2$ ($\rho = 0$).

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.

GEL: Special cases, Exponential Tilting Estimator

$\lambda \longrightarrow -1$ ($\rho = 1$).

The second case is the exponential tilting estimator with $\lambda \rightarrow -1$ (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

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

In general the differences between the estimators within this class is small compared to the differences between GEL and the two-step GMM.

In practice the choice between them is largely driven by computational issues.

The empirical likelihood estimator has likelihood interpretation and the resulting optimality properties for its bias-corrected version (Newey and Smith, 2004).

Imbens, Spady and Johnson (1998) argue for 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).

Newey and Smith, 2004

Formal comparison of expectation of second order term:

$$\widehat{\theta} = \theta_0 + \frac{A}{\sqrt{N}} + \frac{B}{N} + o_p\left(N^{-1}\right)$$

Focus on Bias $\mathbb{E}[B]$.

$$\text{Bias}\left(\widehat{\theta}_{GEL}\right) = B_I + \frac{\lambda}{2} \cdot B_\Omega \qquad \text{(recall : } \lambda_{\text{EL}} = 0\text{)}$$

$$\text{Bias}\left(\widehat{\theta}_{GMM}\right) = B_I + B_\Omega + B_G + B_W$$

B_Ω comes from third moment of moments: $\mathbb{E}[\psi^3]$

B_W comes from estimating weight matrix.

B_G comes from correlation between $\partial\psi/\partial\theta$ and ψ.

Testing

Likelihood Ratio test:

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

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

where $\hat{\Delta}$ is some estimate of the covariance matrix of the moments.

Lagrange Multiplier test, based on estimated lagrange multipliers $\hat{t}$

$$LM = \hat{t}' \hat{\Delta} \hat{t}.$$

4. Computational Issues

In principle the EL estimator has many parameters (π_i and θ), which could lead to computational difficulties.

Solving the First Order Conditions the first order conditions does not work well.

Imbens, Spady and Johnson suggest penalty function approaches which work better, but not great.

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. This appears to work well.

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 for given θ it is easy to numerically solve for t. E.g., in the exponential tilting case, solve

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

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

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

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

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.

The function $t(\theta)$ has analytic derivatives with respect to θ equal to:

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

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

5. A Dynamic Panel Data Model

To get a sense of the finite sample properties of the empirical likelihood estimators we compare two-step GMM and one of the EL estimators (exponential tilting) in the context of a panel data model

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 $i = 1, \ldots, N$.

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 $(T-1)\cdot(T-2)/2$ moments.

Additional $T-2$ moments:

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

Note that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. So, potentially substantial difference between estimators.

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.

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 use $\theta = 0.5$ and $\theta = 0.9$ in the simulations. The first is comparable to the value estimated from the Abowd-Card data.

$\theta = 0.5$	Number of time periods					
	3	4	6	7	9	11
Two-Step GMM						
median bias	-0.00	0.00	-0.00	-0.00	0.00	0.00
relative median bias	-0.07	0.01	-0.06	-0.08	0.09	0.14
median absolute error	0.05	0.03	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.91	0.88	0.91	0.91	0.89	0.90
covarage rate 95% ci	0.95	0.94	0.95	0.96	0.95	0.94
Exponential Tilting						
median bias	-0.00	-0.00	-0.00	-0.00	0.00	0.00
relative median bias	-0.04	-0.02	-0.09	-0.07	0.02	0.10
median absolute error	0.05	0.03	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.90	0.87	0.90	0.92	0.90	0.91
covarage rate 95% ci	0.95	0.94	0.96	0.95	0.95	0.95

$\theta = 0.9$	Number of time periods					
	3	4	6	7	9	11
Two-Step GMM						
median bias	-0.00	0.00	0.00	0.00	0.00	0.00
relative median bias	-0.02	0.08	0.08	0.03	0.08	0.11
median absolute error	0.04	0.03	0.02	0.02	0.01	0.01
coverage rate 90% ci	0.88	0.85	0.80	0.80	0.78	0.76
covarage rate 95% ci	0.92	0.91	0.87	0.85	0.86	0.84
Exponential Tilting						
median bias	0.00	0.00	-0.00	0.00	-0.00	0.00
relative median bias	0.04	0.09	-0.00	0.01	-0.02	0.13
median absolute error	0.05	0.03	0.02	0.02	0.01	0.01
coverage rate 90% ci	0.87	0.86	0.86	0.88	0.87	0.87
covarage rate 95% ci	0.91	0.90	0.91	0.93	0.91	0.93