EC241a **Econometric Theory**

Spring 2004 **UC Berkeley Department of Economics**

Generalized Method of Moments Estimation IV: Empirical Likelihood

In the GMM case we use a very similar idea to demonstrate semi-parametric efficiency, due to Chamberlain (1987). We first discuss a simple just-identified example. Let $X_1, X_2, \ldots$ be iid and satisfy

$$\mathbb{E}\psi(X,\theta) = \mathbb{E}(X-\theta) = 0,$$

for some value θ_0. We again embed the semiparametric model in a fully parametric model. This time it is not quite the true model, but a very close approximation. Suppose X is discrete with known support $\lambda_1, \ldots, \lambda_K$ and with unknown probabilities $\pi_k = Pr(X = \lambda_k)$. Now we have a fully parametric model. What is the variance for the mle for $\theta = \mathbb{E}(X)$? The mle for π_k is $\hat{\pi}_k = \sum_{i=1}^N 1\{x_i = \lambda_k\}/N$. Then the mle for $\theta = \theta(\pi_1, \pi_2, \ldots, \pi_K) = \sum_{k=1}^K \pi_k \lambda_k$ is

$$\hat{\theta}_{mle} = \theta(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_K) = \sum_{k=1}^K \hat{\pi}_k \lambda_k = \bar{x}.$$

Hence the mle for θ is $\bar{x}$ with variance $\mathrm{Var}(X)$. Clearly $\bar{x}$ is also the gmm estimator, and by its equality with the mle estimator for a parametric model it is semiparametrically efficient.

Chamberlain extends this argument to the general GMM case. With some additional effort you can extend the results to the continuous case, but if the estimators are efficient for all discrete distributions, it would seem to be pretty good already because most distributions can be approximated arbitrarily closely by discrete distributions.

In the empirical likelihood argument we take Chamberlain's discrete likelihood argument one step further. Recall that in the just–identified example not only was the variance of the gmm estimator equal to the variance of the discrete–support–mle, in fact the gmm estimator was equal to the mle. Now consider the following over–identified example. The sequence of pairs $(X_1, Y_1), (X_2, Y_2), \ldots$ are iid. The moment functions are:

$$\psi(x, y, \theta) = \begin{pmatrix} \theta - x \\ y \end{pmatrix}.$$

Now assume that (X, Y) is discrete with known support $(\gamma_{x1}, \gamma_{y1}), \ldots, (\gamma_{xK}, \gamma_{yK})$ and

$$\pi_k = Pr(X = \gamma_{xk}, Y = \gamma_{yk}).$$

To estimate the π's we have to take into account the restriction

$$\mathbb{E}(Y) = \sum_{k=1}^{K} \pi_k \gamma_{yk} = 0.$$

Hence we maximize the log likelihood

$$L(\pi) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} 1\{x_i = \gamma_{xk}, y_i = \gamma_{yk}\} \pi_k \right),$$

subject to the restrictions

$$\sum_{k=1}^{K} \pi_k \gamma_{yk} = 0 \qquad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1,$$

over π and then solve for $\theta = \sum_k \pi_k \gamma_{xk}$. (An aside: alternatively we can add an additional restriction that $\sum_k \pi_k(\theta - \gamma_{xk}) = 0$, and maximize over both π and θ. That would give the same numerical results. The Lagrange multiplier for this additional restriction would be equal to zero.)

Let $\delta_{ik} = 1\{x_i = \gamma_{xk}, y_i = \gamma_{yk}\}$, and let $\hat{p}_k = \sum_{i=1}^{N} \delta_{ik}/N$ be the sample frequency for the kth support point. The first order conditions are

$$\sum_{i=1}^{N} \frac{\delta_{ik}}{\pi_k} - \lambda \cdot \gamma_{yk} - \mu,$$

for $k = 1, \ldots, K$, and

$$\hat{\theta} = \sum_{k=1}^{K} \pi_k \gamma_{xk},$$

with λ the Lagrange multiplier for the restriction $\sum \pi\gamma_{yk} = 0$ and μ the Lagrange multiplier for the restriction that $\sum \pi_k = 1$.

Multiply the first order condition by π_k and add up over $k = 1$ to K to get:

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_k - \lambda\sum_{k=1}^{K}\gamma_{yk}\cdot\pi_k - \mu\sum_{k=1}^{K}\pi_k = N - \mu.$$

Hence $\mu = N$. This implies

$$\pi_k = \hat{p}_k/(1 + \gamma_{yk}\cdot\lambda/N).$$

Hence the solution for λ can be characterized through the equation

$$0 = \sum_{k=1}^{K}\pi_k\gamma_{yk} = \sum_{k=1}^{K}\hat{p}_k\gamma_{yk}/(1 + \gamma_{yk}\lambda/N)$$

$$= \sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\delta_{ik}\gamma_{yk}/(1 + \gamma_{yk}\lambda/N) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}\gamma_{yk}/(1 + \gamma_{yk}\lambda/N).$$

Note that summing up over all k the indicator δ_{ik} just picks out the value at the observed value of x or y, so that $\sum_{k=1}^{K}\delta_{ik}\gamma_{yk} = y_i$, and hence we have

$$0 = \frac{1}{N}\sum_{i=1}^{N} y_i/(1 + y_i\lambda/N).$$

Also,

$$\theta = \sum_{k=1}^{K}\pi_k\gamma_{xk} = \sum_{k=1}^{K}\hat{p}_k\gamma_{xk}/(1 + \gamma_{yk}\lambda/N)$$

$$= \sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\delta_{ik}\gamma_{xk}/(1 + \gamma_{yk}\lambda/N)$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i/(1 + y_i\lambda/N).$$

Therefore we can characterize the solution for θ and $t = \lambda/N$ as the solution to the set of equations

$$0 = \sum_{i=1}^{N}\psi(y_i, x_i, t, \theta),$$

with

$$\psi(y,x,t,\theta) = \begin{pmatrix} (\theta - x)/(1+ty) \\ y/(1+ty) \end{pmatrix},$$

where we use the fact that $\sum_{i=1}^N 1/(1+y_i\lambda/N) = \sum_{i=1}^N \pi_k = 1$. This solution is the exact maximum likelihood estimator for θ if the distribution of (Y,X) is discrete. What if the distribution is not discrete? The estimator can still be calculated. Nothing in its formulation requires knowledge of the support of (X,Y). What are its properties in the general case?

First consider the expectation of the moments at $\theta = \theta^* = \mathbb{E}[X]$ and $t = 0$. In that case the two moments have expectation zero. Hence, under the standard regularity conditions the GMM estimator is consistent for $(\theta^*, 0)$. What about the variance? The large sample variance is

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

First the matrix of derivatives:

$$\Gamma = \mathbb{E}\begin{pmatrix} \frac{\partial\psi_1}{\partial\theta}(Y,X,\theta^*,0) & \frac{\partial\psi_1}{\partial t}(Y,X,\theta^*,0) \\ \frac{\partial\psi_2}{\partial\theta}(Y,X,\theta^*,0) & \frac{\partial\psi_2}{\partial t}(Y,X,\theta^*,0) \end{pmatrix} = \begin{pmatrix} 1 & \sigma_{XY} \\ 0 & -\sigma_{YY} \end{pmatrix},$$

Second, the matrix of expected values of the outer products:

$$\Delta = \mathbb{E}\Big[\psi(Y,X,\theta^*,0)\psi(Y,X,\theta^*,0)\Big] = \begin{pmatrix} \sigma_{XX} & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_{YY} \end{pmatrix},$$

with inverse

$$\Delta^{-1} = \frac{1}{\sigma_{XX}\sigma_{YY} - \sigma_{XY}^2} \cdot \begin{pmatrix} \sigma_{YY} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{XX} \end{pmatrix},$$

Hence the covariance matrix is

$$\left(\begin{pmatrix} 1 & 0 \\ \sigma_{XY} & -\sigma_{YY} \end{pmatrix} \cdot \frac{1}{\sigma_{XX}\sigma_{YY} - \sigma_{XY}^2} \cdot \begin{pmatrix} \sigma_{YY} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{XX} \end{pmatrix} \cdot \begin{pmatrix} 1 & \sigma_{XY} \\ 0 & -\sigma_{YY} \end{pmatrix}\right)^{-1}$$

$$= \left(\frac{1}{\sigma_{XX}\sigma_{YY} - \sigma_{XY}^2} \cdot \begin{pmatrix} \sigma_{YY} & 0 \\ 0 & \sigma_{YY}(\sigma_{YY}\sigma_{XX} - \sigma_{XY}^2) \end{pmatrix}\right)^{-1}$$

$$= \begin{pmatrix} \sigma_{XX} - \sigma_{XY}^2/\sigma_{YY} & 0 \\ 0 & 1/\sigma_{YY} \end{pmatrix}.$$

Now compare the variance for $\hat{\theta}$ to the efficiency bound for this case, which corresponds to the variance of the GMM estimator for θ with moments

$$\tilde{\psi}(Y, X, \theta) = \begin{pmatrix} \theta - X \\ Y \end{pmatrix}.$$

This variance would be

$$\left(\begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \frac{1}{\sigma_{XX}\sigma_{YY} - \sigma_{XY}^2} \cdot \begin{pmatrix} \sigma_{YY} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{XX} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right)^{-1} = \sigma_{XX} - \sigma_{XY}^2/\sigma_{YY},$$

which is the same. Hence this <u>empirical likelihood</u> estimator is efficient.

In the previous example, with the moment functions,

$$\psi(x, y, \theta) = \begin{pmatrix} \theta - x \\ y \end{pmatrix},$$

we found for the asymptotic distribution of $\hat{\theta}$ and t:

$$\sqrt{N}(\begin{pmatrix} \hat{\theta} - \theta^* \\ t \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{XX} - \sigma_{XY}^2/\sigma_{YY} & 0 \\ 0 & 1/\sigma_{YY} \end{pmatrix} \right).$$

Note that the estimator for $t = 0$ is independent of the estimator for θ. This makes sense: if there was some correlation there we could improve our estimate of θ by exploiting this correlation. The role of t is that of a <u>Lagrange Multiplier</u>. It measures the deviation of the restriction from its restricted value. Specifically we can use it to test the overidentifying restrictions. Recall that in the standard GMM approach we test the overidentifying restrictions by looking at

$$\frac{1}{N} \cdot Q_{C_N}(\hat{\theta}) = \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \hat{\theta}) \right]' \cdot C_N \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \hat{\theta}) \right] \xrightarrow{d} \mathcal{X}^2(M - K).$$

A similar test can be based on $\sqrt{N} \cdot t$:

$$\frac{1}{N} \cdot Q_{C_N}(\hat{\theta}) - N \cdot t'\Delta \cdot t \xrightarrow{p} 0.$$

Imbens, Spady and Johnson (1997) discuss some of these tests in more detail.

What are the advantages of these empirical likelihood estimators relative to the standard GMM estimators? The main advantage is that there is no ambiguity: there is no need for an initial estimator to start the process, and the estimators are therefore invariant to renormalizations of the moment functions and reparametrizations. It also clarifies Chamberlain's efficiency argument: not only is the variance equal to that of the maximum likelihood estimator under discreteness, we can directly use the maximum likelihood estimator. The general case for the empirical likelihood estimator goes as follows. Suppose we start of with a GMM estimator with moments $\psi(z,\theta)$. We can then get the empirical likelihood estimator as

$$\max_{\theta,\pi} \sum_{i=1}^{N} \ln \pi_i,$$

subject to

$$\sum_{i=1}^{N} \pi_i = 1, \qquad \text{and} \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i,\theta) = 0.$$

This can be rewritten as the GMM estimator with moments

$$\rho(z,t,\theta) = \begin{pmatrix} \psi(z,\theta)/(1+t'\psi(z,\theta)) \\ \frac{\partial \psi'}{\partial \theta}(z,\theta)t/(1+t'\psi(z,\theta)) \end{pmatrix}.$$

Finally, there are other estimators like the empirical likelihood estimator. Another leading member is the estimator based on the Kullback–Leibler information criterion. For the KLIC criterion the general case leads to the objective function

$$\max_{\theta,\pi} -\sum_{i=1}^{N} \pi_i \ln \pi_i,$$

subject to

$$\sum_{i=1}^{N} \pi_i = 1, \qquad \text{and} \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i,\theta) = 0,$$

and the estimating equations

$$\rho(z,t,\theta) = \begin{pmatrix} \psi(z,\theta) \cdot \exp(t'\psi(z,\theta)) \\ \frac{\partial \psi'}{\partial \theta}(z,\theta)t \cdot \exp(t'\psi(z,\theta)) \end{pmatrix}.$$

For the example under consideration this criterion leads to an estimator with moments

$$\psi(y,x,t,\theta) = \begin{pmatrix} (\theta - x) \cdot \exp(ty) \\ y \cdot \exp(ty) \end{pmatrix}.$$

It may be difficult to directly solve the estimating equations. Instead consider the saddlepoint representation:

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

The first order conditions for this problem are the estimating equations. This is still a difficult problem to solve, but an easier computational problem is to concentrate out the Lagrange multipliers at each step. The problem of solving

$$\min_t \sum_{i=1}^{N} \exp(t'\psi(z_i,\theta)), \tag{2}$$

is relatively straightforward, because as a function of t this is strictly convex, with positive definite second derivative

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

Hence the approach is to solve

$$\max_\theta \sum_{i=1}^{N} \exp(t(\theta)'\psi(z_i,\theta)), \tag{3}$$

where

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

Finally consider the following special case (Imbens and Hellerstein) where some of the moments do not depend on the unknown parameters:

$$\psi(z,\theta) = \begin{pmatrix} \psi_1(z,\theta) \\ \psi_2(z) \end{pmatrix}.$$

The first set of moments is of dimension K, equal to the dimension of θ, and the second is of dimension $M-K$, but does not depend on θ. In that case things simplify considerably. Starting with

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

we obtain in general the estimating equations

$$\rho(z,t,\theta) = \begin{pmatrix} \psi(z,\theta)\cdot\exp(t'\psi(z,\theta)) \\ t'\frac{\partial\psi}{\partial\theta}(z,\theta)\cdot\exp(t\psi(z,\theta)) \end{pmatrix},$$

which, after splitting t into t_1 and t_2 with dimension K and $M-K$, can be written as

$$\rho(z,t_1,t_2,\theta) = \begin{pmatrix} \psi_1(z,\theta)\cdot\exp(t_1'\psi_1(z,\theta)+t_2'\psi_2(z)) \\ \psi_2(z)\cdot\exp(t_1'\psi_1(z,\theta)+t_2'\psi_2(z)) \\ \left(\frac{\partial\psi_1'}{\partial\theta}(z,\theta)t_1+\frac{\partial\psi_2'}{\partial\theta}(z,\theta)t_2\right)\cdot\exp(t_1'\psi_1(z,\theta)+t_2'\psi_2(z)) \end{pmatrix}.$$

Because $\frac{\partial\psi_2'}{\partial\theta}(z,\theta)=0$, this simplifies to

$$\rho(z,t_1,t_2,\theta) = \begin{pmatrix} \psi_1(z,\theta)\cdot\exp(t_1'\psi_1(z,\theta)+t_2'\psi_2(z)) \\ \psi_2(z)\cdot\exp(t_1'\psi_1(z,\theta)+t_2'\psi_2(z)) \\ \frac{\partial\psi_1'}{\partial\theta}(z,\theta)t_1\cdot\exp(t_1'\psi_1(z,\theta)+t_2'\psi_2(z)) \end{pmatrix}.$$

Typically a solution can be found with $t_1=0$, and t_2 and $\hat\theta$ solving

$$0=\sum_{i=1}^{N} = \begin{pmatrix} \psi_1(z,\theta)\cdot\exp(t_2'\psi_2(z)) \\ \psi_2(z)\cdot\exp(t_2'\psi_2(z)) \end{pmatrix}.$$

Now we can first find the value of t_2 that solves

$$\sum_{i=1}^{n}\psi_2(z_i)\exp(t_2'\psi_2(z_i))=0.$$

Given this value of t, get $\hat\theta$ by solving the weighted set of moments:

$$\sum_{i=1}^{n}\psi_1(z_i,\theta)\cdot w_i=0.$$

where

$$w_i=\exp(t'\psi_2(z_i))\Big/\sum_{j=1}^{n}\exp(t'\psi_2(z_j)).$$

Additional References

Imbens, Spady, and Johnson, (1998), "Information–theoretic Approaches to Inference in Moment Condition Models", *Econometrica*, Vol. 66, 333-357.

Qin, and J. Lawless, (1994), "Generalized Estimating Equations" *Annals of Statistics*, Vol. 22, 300-325.