GMM Notes for EC2610

1 Introduction

These notes povide an introduction to GMM estimation. Their primary purpose is to make the reader familiar enough with GMM to be able to solve problem set assignments. For the more theoretical foundations, properties and extensions of GMM, or to better understand its workings, interested reader should consult any of the standard graduate econometrics textbooks, e.g., by Greene, Wooldridge, Hayashi, Hamilton, etc., as well as the original GMM article by Hansen (1982). Available lecture notes for graduate econometrics courses, e.g. by Chamberlain (Ec 2140), by Pakes and Porter (Ec 2144), also contain very useful reviews of GMM.

Generalized Method of Moments provides asymptotic properties for estimators and is general enough to include many other commonly used techniques, like OLS and ML. Having such an umbrella to encompass many of the estimators is very useful, as one doesn't have to derive each estimator property separately. With such a wide range, it is not surprising to see GMM used extensively, but one should also be careful when it is appropriate to apply. Since GMM deals with asymptotic properties, it works well for large samples, but does not provide an answer when the samply size is small, or what is "large" enough sample size. Also, when applying GMM, one may forgo certain desirable properties, like efiiciency.

2 GMM Framework

2.1 Definition of GMM Estimator

Let x_i, $i = 1, ..., n$ be i.i.d. random draws from the unknown population distribution P. For a known function ψ, the parameter $\theta_0 \in \Theta$ (usually also in the interior of Θ) is known to satisfy the *key moment condition:*

$$E\left[\psi(x_i, \theta_0)\right] = 0 \tag{1}$$

This equation provides the core of the GMM estimation. The appropriate function ψ and the parameter θ_0 are usually derived from a theoretical model. Both ψ and θ_0 can be vector valued and not necessarily of the same size. Let the size of ψ be q, and the size of θ be p. The mean is 0 only at the true parameter value θ_0, which is assumed to be unique over some neighborhood around θ_0. Along with equation (1), one also imposes certain boundary conditions for the 2nd order moment and partial derivative one:

$$E\left[\psi(x_i, \theta_0)\psi'(x_i, \theta_0)\right] \equiv \Phi < \infty$$

and

$$\left|\frac{\partial^2 \psi_j(x, \theta)}{\partial \theta_k \partial \theta_l}\right| \leq m(x)$$

for all $\theta \in \Theta$, where $E[m(x)] < \infty$. Also, define

$$D \equiv E\left[\frac{\partial \psi(x_i, \theta_0)}{\partial \theta'}\right]$$

and assume, D has rank equal to p, the dimension of θ.

(Note: the above conditions are sufficient, and properties of GMM estimators can also be obtained under weaker conditions).

The task of the econometrician lies in obtaining estimate $\widehat{\theta}$ of θ_0 from the key moment condition. Since there is sample of size n from the population distribution, one may try to obtain the estimate by replacing the population mean with a sample one:

$$\frac{1}{n}\sum_i \psi(x_i, \widehat{\theta}) = 0 \tag{2}$$

This is a system of q equations with p unknowns. If p=q, we're "just-identified," and under some weak conditions, one can obtain a (unique) solution to (2) around the neighborhood of θ_0. When q>p, then we're "over-identified," and a solution will not exist for most functions ψ. A natural approach for the latter case might be to try to get the left hand side as close to 0 as possible, with "closeness" defined over some norm $\|\cdot\|_{A_n}$:

$$\|y\|_{A_n} = y' A_n^{-1} y$$

where A_n is q-by-q symmetric, positive definite matrix.

Another approach could be to find the soltuion to (2), by making some linear combination of ψ_j equations equal to 0. I.e. for some p-by-q matrix C_n, of rank p, solve for:

$$C_n \frac{1}{n}\sum_i \psi(x_i, \widehat{\theta}) = 0 \tag{3}$$

which will give us p equations with p unknowns.

In fact, both approaches are equivalent and GMM estimation is setup to do exactly that. That is, when p=q, GMM is just-identified and we can usually solve for $\widehat{\theta}$ exactly. When q>p, we're in the over-identified case and for some appropriate matrix A_n (or C_n), GMM estimate $\widehat{\theta}$ is found by:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \left[\frac{1}{n}\sum_i \psi(x_i, \theta)\right]' A_n^{-1} \left[\frac{1}{n}\sum_i \psi(x_i, \theta)\right] \tag{4}$$

(Or equivalently, solving for:equation (3)). The choice of A_n will be discussed later, but for now assume $A_n \longrightarrow \Psi$ a.s., where Ψ is also symmetric and positive-definite.

2.2 Asymptotic properties of GMM

Given the above setup, GMM provides two key results: consistency and asymptotic normality. Consistency shows that our estiamtor gives us the "right"

answer, and asymptotic normality provides us with variance-covariance matrix, which we can use for hypothesis testing. More specifically, the estimator $\widehat{\theta}$, found via equation (3) satisfies $\widehat{\theta} \longrightarrow \theta_0$ a.s. (consistency), and

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Lambda) \tag{5}$$

(asymptotic normality), where

$$\Lambda = \Delta\Phi\Delta'$$

and

$$\Delta = (D'\Psi^{-1}D)^{-1}D'\Psi^{-1}$$

(Looking at above properties, one can draw obvious similarities between the GMM estimator, and the Delta Method).

To do hypothesis testing, let $\overset{A}{\sim}$ denote the asymptotic distribtuion. Then, equation (5) implies:

$$\widehat{\theta} \overset{A}{\sim} N(\theta_0, \frac{1}{n}\Lambda)$$

where

$$\begin{aligned} \Lambda &= \Delta\Phi\Delta' \\ &= (D'\Psi^{-1}D)^{-1}D'\Psi^{-1}\Phi\Psi^{-1}D(D'\Psi^{-1}D)^{-1} \end{aligned} \tag{6}$$

Φ and D are population means defined over true parameter values, and Ψ is the probability limit of A_{n}..When computing the variance matrix for a given sample, one usually replaces the population mean with the sample mean; the true parameter value with the estimated value, and Ψ with A_n :

$$\begin{aligned} \Phi &= E\left[\psi(x_i, \theta_0)\psi'(x_i, \theta_0)\right] \\ &\approx \frac{1}{n}\sum_i \psi(x_i, \widehat{\theta})\psi'(x_i, \widehat{\theta}) \end{aligned}$$

$$\begin{aligned} D &= E\left[\frac{\partial\psi(x_i, \theta_0)}{\partial\theta'}\right] \\ &\approx \frac{1}{n}\sum_i \frac{\partial\psi(x_i, \theta)}{\partial\theta'} \mid_{\theta=\widehat{\theta}} \end{aligned}$$

and

$$\Psi \approx A_n$$

The standard errors are obtained from:

$$SE_k = \sqrt{\frac{1}{n}\Lambda_{kk}}$$

where Λ_{kk} is the kth diagonal entry of Λ.

2.3 Optimal Weighting Matrices

2.3.1 Choice of $\mathbf{A}_n$

Having established the properties of GMM, we now turn to the choice of the weighting matrix A_n and C_n. When GMM is just identified, then one can usually solve for $\widehat{\theta}$ from equation (2). This is equivalent for finding a unique minimum point in equation (4) for *any* positive-definite matrix A_n. Also, D will be square; and since it has full rank, will be invertible. Then, the variance matrix will be:

$$\begin{aligned}\Lambda &= \Delta\Phi\Delta' \\ &= (D'\Psi^{-1}D)^{-1}D'\Psi^{-1}\Phi\Psi^{-1}D(D'\Psi^{-1}D)^{-1} \\ &= D^{-1}\Psi D'^{-1}D'\Psi^{-1}\Phi\Psi^{-1}DD^{-1}\Psi D'^{-1} \\ &= D^{-1}\Phi D'^{-1}\end{aligned}$$

As expected, the choice of A_n doesn't affect the asymptotic distribution for the just-identified case.

For the over-identified case, the choice of the weight matrix will now matter for $\widehat{\theta}$. However, since the consistency and asymptotic normality results of GMM do not depend on the choice of A_n (as long as it's symmetric and positive definite), we should get our main results again for *any* choice of A_n. In such a case, the most common choice is the identity matrix:

$$A_n = I_q$$

Then, $\Psi = I_q$ and

$$\begin{aligned}\Delta &= (D'\Psi^{-1}D)^{-1}D'\Psi^{-1} \\ &= (D'D)^{-1}D'\end{aligned}$$

and the approximate variance-covariance matrix will be:

$$\begin{aligned}\frac{1}{n}\Lambda &= \frac{1}{n}\Delta\Phi\Delta' \\ &= \frac{1}{n}(D'D)^{-1}D'\Phi D(D'D)^{-1}\end{aligned}$$

(This is the format of GMM variance-covariance matrix Prof. Pakes uses in the IO lecture notes.)

Given that one is free to choose which particular A_n to choose, one can try pick the weighting matrix to give GMM other desirable properties as well, like efficiency. From equation 6, we know that:

$$\Lambda = (D'\Psi^{-1}D)^{-1}D'\Psi^{-1}\Phi\Psi^{-1}D(D'\Psi^{-1}D)^{-1}$$

Since we're now free to pick Ψ, one can choose it to minimzes the variance:

$$\begin{aligned}\Psi^* &= \arg\min_{\Psi} \Lambda \\ &= \arg\min_{\Psi} (D'\Psi^{-1}D)^{-1}D'\Psi^{-1}\Phi\Psi^{-1}D(D'\Psi^{-1}D)^{-1}\end{aligned}$$

It is easy to show that the minimum is equal to:

$$\min_{\Psi} (D'\Psi^{-1}D)^{-1}D'\Psi^{-1}\Phi\Psi^{-1}D(D'\Psi^{-1}D)^{-1} = (D'\Phi^{-1}D)^{-1}$$

which is obtained at

$$\Psi^* = \Phi$$

The above solution has very intuitive appeal: indexes with larger variances are assigned smaller weights in the estimation.

2.3.2 2-Step GMM estimation

The above procedure then gives rise to 2-step GMM estimation, in the spirit of FGLS.

1. Pick $A_n = I$ (equal weighting), and solve for the 1st stage GMM estimate: $\widehat{\theta}_1$. Since $\widehat{\theta}_1$ is consistent, $\frac{1}{n}\sum_i \psi(x_i, \widehat{\theta}_1)\psi(x_i, \widehat{\theta}_1)'$ will be consistent estimate of Φ.

2. Pick $A_n = \frac{1}{n}\sum_i \psi(x_i, \widehat{\theta}_1)\psi(x_i, \widehat{\theta}_1)'$, and obtain the 2nd stage GMM estimate $\widehat{\theta}_2$. The variance matrix $\frac{1}{n}\widehat{\Lambda}_2$ will then be the smallest.

2.3.3 Choice of $\mathbf{C}_n$

It should be clear by now how the equations (3) and (4) are related to each, and correspondingly, how A_n and C_n are related. By actually differentiating the minimization problem in equation (4), we obtain the FOC:

$$\left[\frac{1}{n}\sum_i \frac{\partial\psi(x_i, \widehat{\theta})}{\partial\theta'}\right]' A_n^{-1} \left[\frac{1}{n}\sum_i \psi(x_i, \widehat{\theta})\right] = 0 \tag{7}$$

If we now define

$$C_n \equiv \left[\frac{1}{n}\sum_i \frac{\partial\psi(x_i, \widehat{\theta})}{\partial\theta'}\right]' A_n^{-1}$$

we have equation (7) turning into (3).

One caveat should be pointed out. We specified that equation (3) is linear combination of $\psi_j(x, \widehat{\theta})$, i.e. C_n is a matrix of constants. But in equation (7) C_n will in general depend on the solution of the equation: $\widehat{\theta}$. This can be easily circumvented if we look at the 2nd stage GMM solution, and use the first stage 1st stage $\widehat{\theta}_1$ for C_n. That is, if in the second step, we'd normally solve:

$$\left[\frac{1}{n}\sum_i \frac{\partial\psi(x_i, \widehat{\theta}_2)}{\partial\theta'}\right]' A_n^{-1} \left[\frac{1}{n}\sum_i \psi(x_i, \widehat{\theta}_2)\right] = 0$$

where A_n is obtained from the 1st stage. We can instead solve for a different 2nd stage estimate $\widehat{\theta}_{2'}$:

$$\left[\frac{1}{n}\sum_i \frac{\partial\psi(x_i,\widehat{\theta}_1)}{\partial\theta'}\right]' A_n^{-1}\left[\frac{1}{n}\sum_i \psi(x_i,\widehat{\theta}_{2'})\right] = 0$$

Since $\widehat{\theta}_1$ satisfies consistency, and asymptotic normality, $\widehat{\theta}_{2'}$ will once again be consistent, asymptotically normal, as well as efficient among the class of GMM estimators. And now C_n is linear when solving for $\widehat{\theta}_{2'}$.

3 Applications of GMM

3.1 Ordinary Least Squares

Since GMM does not impose any restrictions on the functional form of ψ, it can be easily applied to simple-linear as well as non-linear moment conditions. (It can also be extended to continuous, but non-differentiable functions). The usefulness of GMM is perhaps more evident for non-linear estimations, but one can become more familiar with GMM by drawing similarities with other standard techniques.

For the case of OLS, we have:

$$y_i = x_i'\beta + \varepsilon_i$$

with the zero covariance condition:

$$E(x_i\varepsilon_i) = 0$$

The latter is the key GMM moment condition, and can be rewritten as:

$$\begin{aligned} E(\psi(x_i,\beta)) &= 0 \\ E(x_i(y_i - x_i'\beta)) &= 0 \end{aligned}$$

The sample analog becomes:

$$\frac{1}{n}\sum_i (x_i(y_i - x_i'\widehat{\beta})) = 0$$

Since these are k equations with k unknowns, GMM is just-identified with the unique solution of:

$$\widehat{\beta}_{GMM} = \left(\sum_i x_i x_i'\right)^{-1}\left(\sum_i x_i y_i\right)$$

which corresponds to the OLS solution. For the variance covariance matrix we need to compute only Φ and D :

$$\begin{aligned}
\Phi &= E(\psi(x_i,\beta)\psi(x_i,\beta)') \\
&= E(x_i\varepsilon_i\varepsilon_i x_i') \\
&= E(\varepsilon_i^2 x_i x_i') \\
\widehat{\Phi} &= \frac{1}{n}\sum_i e_i^2 x_i x_i'
\end{aligned}$$

where $e_i = y_i - x_i'\widehat{\beta}$.

$$\begin{aligned}
D &= E\left[\frac{\partial\psi(x_i,\beta_0)}{\partial\beta'}\right] \\
&= E\left(-x_i x_i'\right) \\
\widehat{D} &= -\frac{1}{n}\sum_i x_i x_i'
\end{aligned}$$

Then, the variance-covariance matrix will equal:

$$\begin{aligned}
\frac{1}{n}\widehat{\Lambda} &= \frac{1}{n}\widehat{D}^{-1}\widehat{\Phi}\widehat{D}'^{-1} \\
&= \frac{1}{n}\left(-\frac{1}{n}\sum_i x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_i e_i^2 x_i x_i'\right)\left(-\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \\
&= \left(\sum_i x_i x_i'\right)^{-1}\left(\sum_i e_i^2 x_i x_i'\right)\left(\sum_i x_i x_i'\right)^{-1}
\end{aligned}$$

This is also known as the White formula for heteroskedasticity-consistent standard errors.

For a simpler OLS example, if we assume homoskedasticity, one can also obtain a simpler version of the variance matrix. With homoskedasticity,

$$\begin{aligned}
\Phi &= E(\varepsilon_i^2 x_i x_i') \\
&= E(E(\varepsilon_i^2 \mid x_i) x_i x_i') \\
&= E\left(E(\varepsilon_i^2) x_i x_i'\right) \\
&= E(\varepsilon_i^2)E(x_i x_i') \\
\widehat{\Phi} &= \left(\frac{1}{n}\sum_i e_i^2\right)\left(\frac{1}{n}\sum_i x_i x_i'\right)
\end{aligned}$$

Then,

$$\frac{1}{n}\widehat{\Lambda} = \left(\frac{1}{n}\sum_i e_i^2\right)\left(\sum_i x_i x_i'\right)^{-1}$$

which is the variance estimate for the homoskedastic case.

3.2 Instrumental Variables

Suppose again

$$y_i = x_i'\beta + \varepsilon_i$$

But for the IV estimation, we have

$$E(w_i\varepsilon_i) = 0$$

where w_i is not necessarily equal to x_i. We only require $E(w_i'x_i) \neq 0$ to be able to invert matrices. The sample analog now becomes:

$$\frac{1}{n}\sum_i w_i(y_i - x_i'\widehat{\beta}) = 0$$

If w_i and x_i have the same dimension, then we're again in the just-identified case, with the unique solution of:

$$\widehat{\beta}_{GMM} = \left(\sum_i w_i x_i'\right)^{-1}\left(\sum_i w_i y_i\right) = \widehat{\beta}_{IV}$$

If the number of instruments exceeds the number of right-hand side variables, we're in the over-identified case. We can go ahead with 2-stage estimation, but a particular choice of the weighting matrix deserves attention. If we set

$$A_n = \frac{1}{n}\sum_i w_i w_i'$$

The FOC for GMM becomes:

$$\begin{aligned}\left[\frac{1}{n}\sum_i \frac{\partial\psi(x_i,\widehat{\beta})}{\partial\beta'}\right]' A_n^{-1}\left[\frac{1}{n}\sum_i \psi(x_i,\widehat{\beta})\right] &= 0\\ \left(\frac{1}{n}\sum_i w_i x_i'\right)'\left(\frac{1}{n}\sum_i w_i w_i'\right)^{-1}\left(\frac{1}{n}\sum_i w_i(y_i - x_i'\widehat{\beta})\right) &= 0\end{aligned}$$

Let,

$$\widehat{\Pi} = \left(\sum_i w_i w_i'\right)^{-1}\left(\sum_i w_i x_i'\right) \tag{8}$$

$\widehat{\Pi}$ is then the regression coefficients of x_i on w_i.
Then we have:

$$\begin{aligned}\widehat{\Pi}'\left(\sum_i w_i(y_i - x_i'\widehat{\beta})\right) &= 0 \qquad (9)\\ \sum_i \widehat{\Pi}' w_i(y_i - x_i'\widehat{\beta}) &= 0\end{aligned}$$

Note that:

$$\widehat{\Pi}' w_i = \left(w_i' \widehat{\Pi} \right)' = \widehat{x}_i$$

and

$$\sum_i \widehat{x}_i x_i' = \sum_i \widehat{x}_i \widehat{x}_i'$$

where $\widehat{x}_i$ are the fitted values of x_i from (8).
Equation (9) then becomes:

$$\sum_i \widehat{x}_i (y_i - \widehat{x}_i' \widehat{\beta}) = 0$$

This is the solution to 2 Stage Least Squares (2SLS). The first stage is the regression of the right hand side variables on the instruments; and the second stage is the regression of the dependent variable on the fitted values of the right-hand side variables. Thus, with GMM we're able to obtain the 2SLS estimates and their correct standard errors. (The usual setting of 2SLS is to regress only the "problematic" right-hand side variables on the instruments, and then use their fitted values. The right-hand side variables, not correlated with the error term, are part of the instruments, and so their fitted values are equal to themselves. We're then doing the exact same regressions).

3.3 Maximum Likelihood

Suppose now we know the family of distributions, $p(\cdot; \theta)$, where the x_i come from but do not know the true parameter value θ_0. Maximum Likelihood solution to finding θ_0 is:

$$\begin{aligned} \widehat{\theta}_{ML} &= \arg\max_{\theta \in \Theta} p(x_1, ..., x_n \mid \theta) \\ &= \arg\max_{\theta \in \Theta} \prod p(x_i \mid \theta) \end{aligned}$$

The maximum point estimate is invariant to monotonic transformations, and so:

$$\begin{aligned} \widehat{\theta}_{ML} &= \arg\max_{\theta \in \Theta} \log \left(\prod p(x_i \mid \theta) \right) \\ &= \arg\max_{\theta \in \Theta} \frac{1}{n} \sum \log p(x_i \mid \theta) \end{aligned}$$

The FOC becomes:

$$\frac{1}{n} \sum \frac{\partial \log p(x_i \mid \widehat{\theta})}{\partial \theta'} = 0 \tag{10}$$

If we let $\psi(x_i \mid \theta) = \frac{\partial \log p(x_i|\theta)}{\partial \theta'}$, the score function, then equation (10) can serve as the sample analog to a key moment condition of the form:

$$E \left(\frac{\partial \log p(x_i \mid \theta)}{\partial \theta'} \right) = 0$$

When doing ML estimation, the above equation will usually hold. If in doubt, you should consult the references.