Some GMM Notes

Chase Abram*

September 1, 2022

1 Theory

First, we consider the basic theoretical underpinnings of the Generalized Method of Moments (GMM). This section is meant to aid in intuition and conceptual understanding, so the "proofs" are missing details, and the mathematical assumptions are not all specified (i.e. things are differentiable where and when we need them to be, but we say nothing more on that).

Most the below ideas, except for the examples, are essentially my watered down summary of Ali Hortacsu's lecture notes on GMM. They include more detail about the technical assumptions needed.

1.1 Economic Assumptions

We start with an economic model that assumes a structure between the data, X, and parameters, β, which we represent as $\mathbb{E}[F(X,\beta)] = 0$. This is the end of our economic assumptions. Note that this one assumption may be in terms of a scalar or a vector, and may potentially be simple (think OLS, as below) or really complicated (the data could be financial data and the parameters govern the process for returns or something).

1.2 Derivations in Brief

Let the dimension of β be k, and the dimension of $F(X,\beta)$ be r. If $r < k$, we have too many free parameters, and the model is **under-identified**. If $r = k$, the model is exactly **identified**, assuming that the r conditions implied by F are linearly independent. If they are not, we are essentially back in the $r < k$ case. If $r > k$ (and from here on out we assume that r is referring the number of linearly independent conditions given by F), then the model is **over-identified**. Heuristically, when under-identified, we should be able to hit all the moment conditions and even have dimensions(s) of freedom to play with. When identified, we should uniquely hit the moment conditions. When over-identified, we cannot hit all the moment conditions, so what do we do?

Perhaps the most intuitive guess is to throw some moments away. If we let A be the $k \times r$ matrix with a 1 in entry (i,i) for $i \in 1,\dots,k$ and zeros elsewhere, then the following equation is the case where we throw away the last $r-k$ moment conditions.

$$A\mathbb{E}[F(X,\beta)] = 0$$

*Send corrections or comments to abram@uchicago.edu.

This is pretty unsatisfying. Why did we throw those away, and not the first $r-k$? Or some moments in the middle? Or some other linear combination of moments? It feels like there should be an optimal way to do this, throwing away only the least helpful info, and it turns out there is!

Let's first build an estimator. Let $b_{N,A}$ be the estimator we find by imposing that the "empirical moment" is equal to zero, i.e.

$$\begin{aligned} 0 &= A\left(\frac{1}{N}\sum_{i=1}^{N} F(X_i, b_{N,A})\right) \\ &= Ag_N(b_{N,A}) \end{aligned}$$

This estimator will be consistent (I'm not going to provide the needed extra technical assumptions nor prove here), but how efficient is it? Let's define

$$V \equiv \text{var}(F(X,\beta))$$

Now consider the following expansion

$$\begin{aligned} g_N(b_{N,A}) &= g_N(\beta) + D_N(b_{N,A}-\beta) + H(\cdot) \\ D_N &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial F}{\partial \beta}(X_i,\beta) \end{aligned}$$

The H term is second-order and hence goes to zero "quickly" as $b_{N,A} \to \beta$. Using this expression, we can consider the asymptotic distribution of the estimator by taking the following steps (notice that I sloppily throw in the step where $D_N \to D \equiv \mathbb{E}[\frac{\partial}{\partial\beta}F(X,\beta)]$).

$$\begin{aligned} Ag_N(\beta) + AD_N(b_{N,A}-\beta) + AH(\cdot) &= Ag_N(b_{N,A}) \\ &= 0 \\ \Rightarrow Ag_N(\beta) &\approx -AD_N(b_{N,A}-\beta) \\ \Rightarrow \sqrt{N}(b_{N,A}-\beta) &\approx -(AD)^{-1}A\sqrt{N}g_N(\beta) \\ &= -(AD)^{-1}A\frac{1}{\sqrt{N}}\sum_{i=1}^{n}F(X_i,\beta) \\ &\to_d N(0,(AD)^{-1}AVA'(AD)'^{-1}) \end{aligned}$$

Where the last line is by the CLT. What have we gained? Now we know that we can choose any A with rank k, and get a consistent estimator with variance as above. How do we choose A, then?

Remember the Gauss-Markov Theorem (OLS is BLUE)? We can use the same proof approach to show that the optimal choice is $A^* = D'V^{-1}$. Then the variance of the estimator achieves its lower bound (in the positive semi-definite sense): $\left(D'V^{-1}D'\right)^{-1}$.

1.2.1 Sidepoint on Exactly Identified Case

A natural question (that may help intuition) is whether A matters at all when $r = k$. In terms of consistency and efficiency, the answer is no (provided A is non-singular). To see this, let A be any non-singular matrix. Then, from above, the estimator's variance will be

$$\begin{aligned}(AD)^{-1}AVA'(AD)'^{-1} &= D^{-1}A^{-1}AVA'(A')^{-1}(D')^{-1} && (A \text{ is square})\\ &= D^{-1}VD'^{-1}\\ &= (D'V^{-1}D)^{-1}\end{aligned}$$

Thus, it does not matter what A we use, as it always achieves the efficiency bound. It may be tempting to take the above transformations even in the over-identified case, but then some of the terms will not even be well-defined.

So are these A identical in terms of estimation? Well no, we have only considered consistency and efficiency, both of which are asymptotic properties. So different A may perform better in terms of unbiasedness or some other finite-sample property. Actually this is a general critique of GMM by many econometricians: it may perform horribly in finite samples, so is really only useful if we have reason to believe that the asymptotics take over fairly quickly (for fairly small n sample size).

1.3 Computation

1.3.1 Direct

You should immediately say "Okay...but how am I supposed to pick $A = D'V^{-1}$ if I don't know β (which will affect both terms)? I thought the whole point was to find β? Why are you telling me the way to estimate β involves β?!". Fair. The mathematical answer is that we view both D and V as *functions* of $b_{N,A}$. So we solve

$$D(b_{N,A})'V(b_{N,A})^{-1}g_N(b_{N,A}) = 0$$

This is potentially a non-linear equation, and the application will determine how involved/tricky solving it is. But in principle it is solvable numerically. We might also split the solving and iterate until convergence, so something like

(i) Guess $b_{N,A}$

(ii) Find implied D and V

(iii) Use D and V to find new $b_{N,A}$

(iv) Repeat until convergence

1.3.2 Indirect (Quadratic Form Minimization)

The (more popular, I think?) way of implementing GMM is to view our problem as minimizing the quadratic form

$$\min_b g_N(b)'Wg_N(b)$$

where W is an $r \times r$ weighting matrix. In this case, the FONC[1] is

$$\left(\frac{\partial}{\partial b} g_N(b)\right)' W g_N(b) = 0$$

Since the Jacobian term will go to D a.s., this condition becomes

$$D' W g_N(b) = 0$$

Well hey, that means $D'W = A$, and we know optimally $A = D'V^{-1}$, so optimally $W = V^{-1}$. Similarly to the direct approach, we can then calculate an approximation of V as a function of b (this is just like in Azeem's course), and we now have our problem as

$$\min_b g_N(b)'[V(b)]^{-1} g_N(b)$$

Solving this problem will give us an estimate which is both consistent and efficient. Win!

As a final note, consider the difference between the two approaches. In the first case, we ask the computer to solve a system of (potentially non-linear) equations. In the second case, we ask the computer to solve a minimization problem. Which one is easier for the computer? I don't know. It probably depends on the problem at hand (are you seeing the theme?).

In practice this approach is implemented by the "Two-Step" GMM. Here are the steps:

(i) Pretend $W = I$ (or really any invertible matrix), and solve for b by solving the minimization problem

(ii) Use this estimate to construct an empirical estimate of V, then re-solve for b using $W = V^{-1}$

We could iterate on this until everything converges, but in practice one iteration is often enough. In fact, in terms of consistency, if we had an "infinite" amount of data, performing this two step approach once will give the true values of β and V.

2 Examples

We now translate the above ideas into some well-known example estimators. This is meant to show you that GMM is truly a generalization.

Please take all transposes with a grain of salt, though I think they are mostly correct.

2.1 OLS

Our first model is linear, and has uncorrelated homoskedastic errors

$$\begin{aligned} y &= x\beta + \epsilon \\ \mathbb{E}[x'\epsilon] &= 0 \\ \mathbb{E}[\epsilon\epsilon' \mid x] &= \sigma^2 I \end{aligned}$$

[1] Use matrix calculus. The rules can be derived or Googled, as you prefer.

Then we may translate

$$\begin{aligned} X &\equiv (x, y) \\ \beta &\equiv \beta \\ F(X, \beta) &\equiv x'(y - x\beta) \\ \Rightarrow D &\equiv \mathbb{E}[\frac{\partial F}{\partial \beta}(X, \beta)] \\ &= -\mathbb{E}[x'x] \\ V &\equiv \mathbb{E}[(x'\epsilon)(x'\epsilon)'] \\ &= \mathbb{E}[x'\epsilon\epsilon'x] \\ &= \sigma^2\mathbb{E}[x'x] \end{aligned}$$

Then our moment condition gives

$$\begin{aligned} 0 &= \mathbb{E}[F(X, \beta)] \\ &= A\mathbb{E}[x'(y - x\beta)] \\ &= D'V^{-1}\mathbb{E}[x'(y - x\beta)] \\ &= -E[x'x](\sigma^2\mathbb{E}[x'x])^{-1}\mathbb{E}[x'(y - x\beta)] \\ \Rightarrow \mathbb{E}[x'y] &= \mathbb{E}[x'x\beta] \\ \Rightarrow \beta &= \mathbb{E}[x'x]^{-1}\mathbb{E}[x'y] \end{aligned}$$

2.2 OLS (heteroskedastic correction)

Now suppose the errors may be heteroskedastic, so we cannot simplify beyond

$$\begin{aligned} V &\equiv \mathbb{E}[x'\epsilon\epsilon'x] \\ &= \mathbb{E}[x'(y - x\beta)(y - x\beta)'x] \end{aligned}$$

Then the above calculation goes through with a minor tweak

$$\begin{aligned} 0 &= \mathbb{E}[F(X, \beta)] \\ &= A\mathbb{E}[x'(y - x\beta)] \\ &= D'V^{-1}\mathbb{E}[x'(y - x\beta)] \\ &= -E[x'x]\mathbb{E}[x'(y - x\beta)(y - x\beta)'x]^{-1}\mathbb{E}[x'(y - x\beta)] \end{aligned}$$

It now becomes apparent that this "minor tweak" actually makes our lives much more difficult. In practice, we would start with regular OLS, calculate the implied heteroskedasticity correction, then use that to re-estimate, and continue until convergence. It should be noted, however, that instead of developing a whole new theory for solving this heteroskedasticity problem, GMM just let us change one line of math, and recognize our estimator as solving a (potentially non-linear) problem.

2.3 IV/2SLS

Now we assume our error may be correlated with our regressors, but we have an instrument, z, with dimension greater than or equal to x.

$$\begin{aligned} y &= x\beta + \epsilon \\ \mathbb{E}[\epsilon] &= 0 \\ \mathbb{E}[x'\epsilon] &\neq 0 \\ \mathbb{E}[z'\epsilon] &= 0 \\ \mathbb{E}[z'x] &\neq 0 \\ \mathbb{E}[\epsilon\epsilon' \mid z] &= \sigma^2 I \end{aligned}$$

Note that we have assumed the instrument is exogenous and relevant, and again returned to homoskedastic errors. Let's translate

$$\begin{aligned} X &\equiv (x, y, z) \\ \beta &\equiv \beta \\ F(X, \beta) &\equiv z'(y - x\beta) \\ D &\equiv \mathbb{E}[\frac{\partial F}{\partial \beta}(X, \beta)] \\ &= -\mathbb{E}[z'x] \\ V &\equiv \mathbb{E}[(z'\epsilon)(z'\epsilon)'] \\ &= \sigma^2 \mathbb{E}[z'z] \end{aligned}$$

So we find

$$\begin{aligned} 0 &= D'V^{-1}\mathbb{E}[z'(y - x\beta)] \\ &= \mathbb{E}[z'x]'(\sigma^2\mathbb{E}[z'z])^{-1}\mathbb{E}[z'(y - x\beta)] \\ \Rightarrow \beta &= \left(\mathbb{E}[z'x]'\mathbb{E}[z'z]^{-1}\mathbb{E}[z'x]\right)^{-1}\left(\mathbb{E}[z'x]'\mathbb{E}[z'z]^{-1}\mathbb{E}[z'y]\right) \end{aligned}$$

If we stop here, we have the two-stage least squares estimator, and we cannot go further if the dimension of z exceeds x. However, if their dimensions are equal, we are in the instrumental variables case, and the $\mathbb{E}[z'x]$ have proper inverses, so the above expression greatly simplifies into

$$\beta = \mathbb{E}[z'x]^{-1}\mathbb{E}[z'y]$$

2.4 Something Spicier

Consider a "strangers on a train" model, wherein each set of covariates kills the errors for the other model. We start with homoskedastic errors.

$$
\begin{aligned}
y_1 &= x_1\beta_1 + \epsilon_1 \\
y_2 &= x_2\beta_2 + \epsilon_2 \\
\mathbb{E}[\epsilon_i] &= 0 \\
\mathbb{E}[x_1'\epsilon_1] &\neq 0 \\
\mathbb{E}[x_1'\epsilon_2] &= 0 \\
\mathbb{E}[x_2'\epsilon_1] &= 0 \\
\mathbb{E}[x_2'\epsilon_2] &\neq 0 \\
\mathbb{E}[x_1'x_2] &= \mathbb{E}[x_2'x_1]' \neq 0 \\
\mathbb{E}[\epsilon_i\epsilon_j' \mid x_1, x_2] &= \sigma^2 I\mathbf{1}\{i = j\}
\end{aligned}
$$

We translate into our GMM notation

$$
\begin{aligned}
X &\equiv (x_1, y_1, x_2, y_2) \\
\beta &\equiv (\beta_1, \beta_2) \\
F(X, \beta) &\equiv \begin{bmatrix} x_2'(y_1 - x_1\beta_1) \\ x_1'(y_2 - x_2\beta_2) \end{bmatrix} \\
D &\equiv \mathbb{E}\begin{bmatrix} -x_2'x_1 & 0 \\ 0 & -x_1'x_2 \end{bmatrix} \\
V &\equiv \mathbb{E}\begin{bmatrix} x_2'\epsilon_1\epsilon_1'x_2 & 0 \\ 0 & x_1'\epsilon_2\epsilon_2'x_1 \end{bmatrix} \\
&= \sigma^2\begin{bmatrix} x_2'x_2 & 0 \\ 0 & x_1'x_1 \end{bmatrix}
\end{aligned}
$$

Without going through the remaining math, it should be clear at this point (from the fact that D and V are block diagonal) that this resulting estimator is the same as if we just used each set of covariates as an instrument for the other equation. So considering the joint estimator of the whole system does not really buy us anything here, since it the same as estimating each system separately.

Suppose we instead assume

$$
\mathbb{E}[\epsilon_i\epsilon_j' \mid x_1, x_2] = \sigma^2 I
$$

Now the errors within a model do not interact, but they do interact with their corresponding error in the other model.

$$
\begin{aligned}
V &\equiv \mathbb{E}\begin{bmatrix} x_2'\epsilon_1\epsilon_1'x_2 & x_2'\epsilon_1\epsilon_2'x_1 \\ x_1'\epsilon_2\epsilon_1'x_2 & x_1'\epsilon_2\epsilon_2'x_1 \end{bmatrix} \\
&= \sigma^2\begin{bmatrix} x_2'x_2 & x_2'x_1 \\ x_1'x_2 & x_1'x_1 \end{bmatrix}
\end{aligned}
$$

Now to find the estimator we just need some matrix calculus, which is not too difficult, but would not be instructive. The point you should take away, however, is that now, since V is not block-diagonal, **estimating the system jointly will generally be more efficient than estimating**

separately, even though we have the assumptions to do either. So you can choose to estimate the system as two separate equations, and you will still get consistent estimates, but your standard errors will be higher than if you estimated the system jointly, because you are not using the optimal A matrix, whereas the joint system does use $A^* = D'V^{-1}$. Unfortunately, since this system is exactly identified, this point is lost becauase we actually will have that it does not matter what A we use, as long as it is nonsingular. If we had another moment for use in estimation, this point would then be more clear.

2.5 Note on estimation

All the above examples just give an expression for β in terms of moments of distributions. To actually perform the estimation, we would need empirical counterparts, then we need to show they consistently estimate the objects of interest, and to do this we use all the rules about convergence in probability and distributions (e.g. Slutsky's theorem, etc.)