Bent E. Sørensen

November 9, 2022

1 Teaching notes on GMM 1.

NOTE: I rewrote the notes to have consistent notation throughout. (The literature will use different notations, but that cannot well be helped.) Generalized Method of Moment (GMM) estimation is one of two developments in econometrics in the 80ies that revolutionized empirical work in macroeconomics. (The other being the understanding of unit roots and cointegration.)

The path breaking articles on GMM were those of Hansen (1982) and Hansen and Singleton (1982). This paper was so influential that when you/they say "GMM" people often assume/use the estimator in a very specific way which includes a very specific way of estimation variances! (Some people even used to think of it a synonymous with estimation non-linear rational expectations models, but it should be understood by most by now that this is just one very specific application.) In these notes, we will show the specific way that many people use GMM, so you are aware of it. For introductions to GMM, Bruce Hansen's coverage in his Econometrics textbook

is very good. For more comprehensive coverage, see the monograph by Alastair Hall (Oxford University Press 2005).

I think that one can claim that there wasn't that much material in Hansen (1982) that was not already known to specialists, although the article definitely was not redundant, as it unified a large literature (almost every estimator you know can be shown to be a special case of GMM). The demonstration in Hansen and Singleton (1982), that the GMM method allowed for the estimation of non-linear rational expectations models, that could not be estimated by other methods, really catapulted Hansen and Singleton to major fame. We will start by reviewing linear instrumental variables estimation, in a slightly different notation, because that will contain most of the ideas and intuition for the general GMM estimation.

1.1 Linear IV estimation

Consider the following simple model

$$(1) \quad y_t = x_t\theta \ + \ u_t \ , \ t \ = 1, ..., T$$

where y_t and e_t scalar, x_t is $1 \times K$ and θ is a $K \times 1$ vector of parameters. NOTE from the beginning that even though I use the index "t" — indicating time, that GMM methods are applicable, and indeed much used, in cross sectional or panel studies.

In vector form the equation (1) can be written

$$(2)\quad Y = X\theta \ + \ U \ ,$$

in the usual fashion. If x_t and u_t may be correlated, one will obtain a **consistent** estimator by using instrumental variables (IV) estimation. The idea is to find a $1 \times L$ vector z_t that is as highly correlated with x_t as possible and at the same time is independent of u_t—so if x_t is actually uncorrelated with u_t you will use x_t itself as instruments or sometimes a vector on ones—in this way all the simple estimators that you know, like OLS and Maximum Likelihood, are special cases of GMM- estimation. If Z denotes the $T \times L$ $(L \geq K)$ vector of the z-observations then we get by pre-multiplying (2) by Z that

$$(3)\quad Z'Y = Z'X\theta + Z'U \ .$$

If we now denote $Z'Y$ by $\tilde{Y}$, $Z'X$ by $\tilde{X}$, and $Z'E$ by U then the system has the form

$$\tilde{Y} \ = \ \tilde{X}\theta \ + \ M \ ,$$

which corresponds to a standard OLS formulation with L observations. Here the variance Ω of M is

$$\Omega \ = \ var(M) \ = \ Z'var(E)Z \ .$$

Now the standard Least Squares estimator of θ is

$$\hat{\theta} \ = \ (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \ ,$$

which is consistent and unbiased with variance

$$Var(\hat{\theta}) \;=\; (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\Omega\tilde{X}(\tilde{X}'\tilde{X})^{-1} \;.$$

The system (of the form (2)) will often have been obtained via the use of instrumental variables, so minimizing the sum of squares of the *moments* may be IV and not OLS. Most of the GMM-literature uses very sparse notation, which is maybe nice when you are familiar with it, but makes it hard to get started on. (I am not a fan of the sparse notation, sometimes in program package documentations, it can be unclear when they talk about the T-vector $Y - X\beta$ or the moment vector $M = Z'Y - X\beta$.) If M does not have a variance matrix that is proportional to the identity matrix the Least Squares estimator is not efficient. Remember that the Least Squares estimator is chosen to minimize the criterion function

$$M'M \;=\; (\tilde{Y} - \tilde{X}\theta)'(\tilde{Y} - \tilde{X}\theta) \;.$$

To obtain a more **efficient** estimator than the Least Squares estimator we have to give different weights to the different equations. Assume that we have given a **weighting matrix** W (the choice of weighting matrices is an important subject that we will return to) and instead choose $\hat{\theta}$ to minimize

$$M'WM \;=\; (\tilde{Y} - \tilde{X}\theta)'W(\tilde{Y} - \tilde{X}\theta) \;,$$

or (in the typical compact notation)

$$\hat{\theta} \;=\; argmin_{\theta}M'WM \;.$$

In this linear case one can then easily show that $\hat{\theta}$ is the Generalized Least Squares-estimator

$$\hat{\theta} \;=\; (\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\tilde{Y} \;.$$

Let the variance of M be denoted Ω and we find that $\hat{\theta}$ have variance

$$var((\tilde{X}'W\tilde{X})^{-1}\tilde{X}'WM \;=\; (\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\Omega W\tilde{X}(\tilde{X}'W\tilde{X})^{-1} \;.$$

We want to choose the weighting matrix optimally, so as to achieve the lowest variance of the estimator. It is fairly obvious that one will get the most efficient estimator by weighing each equation by the inverse of its standard deviation which suggests choosing the weighting matrix Ω^{-1}. In this case we find by substituting Ω^{-1} for W in the previous equation that

$$var((\tilde{X}'\Omega^{-1}\tilde{X})^{-1}\tilde{X}'\Omega^{-1}M \;=\; (\tilde{X}'\Omega^{-1}\tilde{X})^{-1}\tilde{X}'\Omega^{-1}\Omega\Omega^{-1}X(\tilde{X}'\Omega^{-1}\tilde{X})^{-1} \;=\; (\tilde{X}'\Omega^{-1}\tilde{X})^{-1} \;.$$

We recognize this variance as having the same form of that of the GLS estimator. Since we know that the GLS estimator is the most efficient estimator it must be the case that Ω^{-1} is the optimal weighting matrix. (We usually do GLS on the basic Y and X vectors, but the math is the same, we find coefficients that minimizes the sum of squares.)

For practical purposes one would usually have to do a 2-step estimation. First perform a preliminary estimation by minimizing the sum of squares of the (moment) residuals, then estimate Ω (from the residuals), and perform a second step using this

estimate of Ω to perform "feasible GLS" (still keeping in mind that we are minimizing the moments and not the original error terms). This is asymptotically fully efficient. It sometimes can improve finite sample performance to iterate one step more in order to get a better estimate of the weighting matrix (one may also iterate to joint convergence over Ω and θ — there is some Monte Carlo evidence that this is optimal in small samples).

A special case is the IV estimator (see eq. (3)). If $var(U) = \sigma^2 I$, then the variance of the moments $Z'Y$ is $\sigma^2 Z'Z$. For the weighting, we can ignore the σ^2 and the optimal GMM-estimator is then

$$\hat{\theta} \;=\; (\tilde{X}'(Z'Z)^{-1}\tilde{X})^{-1}\tilde{X}'(Z'Z)^{-1}\tilde{Y} \;,$$

or

$$\hat{\theta} \;=\; (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \;.$$

It is now easy to check that this is the OLS-estimator, when you regress $Z(Z'Z)^{-1}Z'Y$ on $Z(Z'Z)^{-1}Z'X$; i.e., this is the classical IV-estimator, which is referred to as the Two-Stage Least Squares in the context of simultaneous equation estimation. The "first stage" is an OLS-regression on the instrument and the "second stage" is the OLS-regression of the fitted values from the first stage regression.

The derivations above illustrate many of the concepts of GMM. Personally I always

guide my intuition by the GLS model. For the general GMM estimators the formulas look just the same (in particular the formulas for the variance) except that if we consider the nonlinear relation, we would estimate

$$Y = h(X, \theta) \ + \ U \ ,$$

$$(4) \quad Z'Y = Z'h(X, \theta) \ + \ M \ ,$$

then "X" in the GLS-formulas should be changed to $\frac{\partial Z'h}{\partial \theta}$. E.g. using the optimal weighting matrix (much more about that later), you find the *asymptotic* variance of the estimated parameter to be

$$var(\hat{\theta}) = (\frac{\partial Z'h'}{\partial \theta} \Omega^{-1} \frac{\partial Z'h}{\partial \theta})^{-1} \, ,$$

where the derivatives can be thought of as coming from the Delta rule (it is the same underlying logic, asymptotically the estimated parameters converges to the true value and then formulas come from applying first-order Taylor expansions). In GMM jargon, the model would usually be formulated as

$$M \ = \ Z'(Y - h(X, \theta)) \ ,$$

or more often as

$$(**) \quad M(X, \theta) \ .$$

Here, X is redefined to mean all the data series and g is the formula that gives the moment condition. The later—very compact—notation is the one that is commonly

used in the GMM literature and we will follow it here. (In cases, such as the Euler equation, there are no "Y" and "X" variables, but rather an optimality condition involving consumption in different periods. The lack of "dependent" and "independent" variables is one reason that these estimators can seem confusing to some at first.) It is typical for the newer methods (typically inspired from statistics) that the variables are treated symmetrically.

In the language of GMM the whole model is summarized by L **orthogonality conditions**:

$$EM \;=\; 0\;,$$

or (when you want to be really explicit!):

$$EM(X,\theta) \;=\; 0\;.$$

Here you should think of M as being a theoretical model (including justification for the moments) although a theoretical model can be simply assuming X follows a multivariate normal distribution (usually with constraints on the covariances, e.g. i.i.d. or AR(1), etc.). But in the usual formulation of GMM the dimension L of M is fixed, so e.g. in the OLS model where the dimension of $U = \{u_1, ..., u_T\}'$ depends on T, you would think of the orthogonality conditions as being $M = X'Y - X'X\theta$. In rational expectations models, the theory often implies which variables will be valid instruments; but this is not always so. For the statistical development the terse notation is good; but in applications you will of course have to be more explicit.

GMM and Method of Moments

If we have L orthogonality conditions summarized in a vector function $M(X, \theta)$ that satisfies $EM(X, \theta) = 0$, the GMM estimator attempts to minimize a quadratic form in M, namely $M'WM$. Notice that there are L orthogonality conditions (rather than T)—this means that you should think about $Z'(Y - X\theta)$ in the IV setting [rather than $(Y - X\theta)$]. Notice that $Z'(Y - X\theta)$ is a vector and the l'th row is the inner product of the l'th instrument with the time t residual: $\Sigma_{t=1}^{T} Z_{lt}(Y_t - X_t\beta)$. More generally, $Y_t - X_t\beta$ may itself be a vector.

A special case is that Z is just columns of ones—then a relation like $M(X, \theta) = Z'u(X, \theta)$, where u is some "pre-instrument" economic relation, such as the Euler equation, is just $M(X, \theta) = \Sigma_{t=1}^{T} u_t(\theta)$ (where $u_t(\theta) = u(X_t, \theta)$). The notation can be a little confusing, partly because it is so general. One way of getting it straight is to study the program for the homework and make a little flow diagram of what the subroutines do.

When the instrument is a constant, the orthogonality condition is the first empirical moment of the u_t vector. In the case of instruments z_t the orthogonality condition is

$M_T(X,\theta) = \Sigma z_t u_t(X,\theta)$. If the number of orthogonality conditions is the same as the number of parameters you can solve for the θ vector which makes $M_T = 0$—in this case the weighting matrix does not matter. This does not mean that the method is only applicable for first moments, for example you could have

$$u_t = \begin{pmatrix} x_t - \mu \\ \\ x_t^2 - \sigma^2 - \mu^2 \end{pmatrix},$$

which, for a vector of constants as the instruments, corresponds to simple method of moments. More generally, a model often implies that the moments is some non-linear functions of the parameters, and those can then be found by matching the empirical moments with the models implied by the model. (The moments used for the GMM-estimator in Melino-Turnbull (1990) and Ho, Perraudin, and Sørensen (1996) are simply matching of moments). The "Generalized" in GMM comes from the fact that we allow more moments than parameters and that we allow for instruments. Sometimes GMM theory will be discussed as GIVE (Generalized Instrumental Variables Estimation), although this is usually in the case of linear models.

1.2 Hansen and Singleton's 1982 model

This is by now the canonical example.

The model in Hansen and Singleton (1982) is a simple non-linear rational expectations representative agent model for the demand for financial assets. The model is

a simple version of the model of Lucas (1978), and here the model is simplified even more in order to highlight the structure. Note that the considerations below are very typical for implementations of non linear rational expectations models.

We consider an agent that maximize a time-separable von Neumann-Morgenstern utility function over an infinite time horizon. In each period the consumer has to choose between consuming or investing. It is assumed that the consumers utility index is of the constant relative risk aversion (CRRA) type. There is only one consumption good (as in Hansen and Singleton) and one asset (a simplification here). The consumers problem is

$$\text{Max } E_t \left[\sum_{j=0}^{\infty} \beta^j \frac{1}{\gamma} C_{t+j}^{\gamma} \right]$$

$$\text{s.t. } C_{t+j} + \Sigma_i I_{t+j}^i \leq \Sigma_i r_{t+j}^i I_{t+j-1}^i + W_{t+j} \; ; \quad j = 0, 1, .., \infty$$

where E_t is the consumer's expectations at time t and

C_t : Consumption

I_t^i : Investment in (one-period) asset i

W_t : Other Income

r_t^i : Rate of Return on asset i

β : Discount Factor

γ : Parameter of Utility Function

If you knew how C_t and investments were determined this model could be used to find r_t^i (which is why it called an asset pricing model), but here we will consider this optimization problem as if it was part of a larger unknown system. Hansen and Singleton's purpose was to estimate the unknown parameters (β and γ), and to test the model.

The first order conditions (called the "Euler equation") for maximum in the model is that

$$C_t^{\gamma-1} = \beta E_t[C_{t+1}^{\gamma-1} r_{t+1}^i] \,,$$

for each asset i. The model can, in general, not be solved for the optimal consumption path and the major insight of Hansen and Singleton (1982) was that knowledge of the Euler equations are sufficient for estimating the model.

The assumption of rational expectations is critical here - if we assume that the agents expectations at time t (as expressed through E_t corresponds to the true expectations as derived from the probability measure that describes that actual evolution of the variables then the Euler equation(s) (there is one for each asset included) can be used to form the non-instrumented "orthogonality condition(s)"

$$u_t(\theta) \;=\; \beta C_t^{\gamma-1} r_t^i \;-\; C_t^{\gamma-1} \;,$$

or

$$u_t(\theta) \;=\; \beta r_t^i (\frac{C_t}{C_{t-1}})^{\gamma-1} \;-1\;,$$

where $E_{t-1}u_t = 0$. (The second term is preferred because C_t behaves similar to a random walk ("unit root process") which growth consumption does not.) Note that $E_{t-1}u_t = 0$ implies that $Eu_t = 0$ by the "law of iterated expectations", which is all that is needed in order to estimate the parameters by GMM. The fact that the *conditional* expectation of u_t is equal to zero can be quite useful for the purpose of selecting instruments. In the Hansen-Singleton model, we have one orthogonality condition and that is not enough in order to estimate two parameters (more about that shortly), but if we can find two or more independent instrumental variables to use as instruments then we effectively have more than 2 orthogonality conditions.

We denote the agents information set at time t by Ω_{t-1}. (This is just jargon, but you may well encounter it.) Ω_{t-1} will typically be a set of, say, K, previous observations of economic variables $\{z_{1,t-1}, z_{1,t-2}...; z_{2,t-1}, z_{2,t-2}, ...; z_{K,t-1}, ...\}$. (Including $C_{t-1}, I_{t-1}^i, r_{t-1}^i$ among the z's if you want.) Any variable in Ω_{t-1} will be a valid instrument in the sense that

$$E[z_{t-j}u_t(\theta)] \;=\; 0$$

for any z_{t-j} in Ω_{t-1}. Notice that z_{t-1} here denotes any valid instrument at time $t-1$, for example z_{t-1} could be C_{1t-3}—this convention of indexing the instruments will prove quite convenient. The $E[.,.]$ operation can be considered an inner product, so

this equation is really the origin of the term orthogonality conditions. For those of you who want to see how this can be developed rigorously, see the book by Hansen and Sargent (1991).

Take a few seconds to appreciate how elegant it all fits together. Economic theory gives you the first order condition directly, then you need instruments, but again they are delivered by the model. For empirical economists who want to derive estimation equations from economic principles, it does not get any better that this.

Oh, well maybe there is a trade-off. The reason being that instrumental variables estimators are not very efficient if no good instruments are available. The literature on "weak instruments" is evolving rapidly, but it seems that sometimes people using GMM-packages forget that they are doing (maybe non-linear) IV (but if you both have non-linearity and weak instruments, "who knows" what you get).

Hansen-Singleton also estimated the Euler equations using Maximum Likelihood in the paper"Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns", JPE, 93, p 249-265. Here they need to impose enough assumptions imposed that the model can be estimated by Maximum Likelihood. Sometimes (maybe a bit less often than in the past) people will stress that GMM is consistent under very weak assumptions. In most cases—meaning unless the sample sizes are very large—I think that is fools' gold—the asymptotic distributions depends on

asymptotic normality and if thing are far from normal, the limit theorems may not be relevant in short samples. And this is when the instruments are not weak. So, be very skeptical if people present super high t-stats in small samples (and don't do it yourself without further verification, such as Monte Carlo).