Bent E. Sørensen

November 10, 2022

1 Teaching notes on GMM II.

Assume that economic theory gives us the moment conditions

$$EM_T(\theta) = 0\,.$$

where the vector M_T' takes the form $(\Sigma_{t=1}^T z_{1,t}u_t, ..., \Sigma_{t=1}^T z_{L,t}u_t)$ when you have only one u variable, and if you have two u-variables u_1 and u_2, like the Euler equations for two different assets, $M_T' = (\Sigma_{t=1}^T z_{1,t}u_{1,t}, ..., \Sigma_{t=1}^T z_{L/2,t}u_{1,t}$, $\Sigma_{t=1}^T z_{1,t}u_{2,t}, ..., \Sigma_{t=1}^T z_{L/2,t}u_{2,t})$. (Note that if we have L moment conditions and 2 basic u variables, then there can only be $L/2$ instruments.) This way is common, but you are allowed to use different instruments and different numbers of instruments for each u, the important thing is you have $K \leq L$ moment conditions so that your parameters are identified. We define the

$L \times 1$ vector $f_t(\theta)$, such that $f_t' = (z_{1,t}u_t, ..., z_{L,t}u_t)'$ (for the case with one u and correspondingly for other cases) such that $M_T(\theta) = \Sigma_{t=1}^T f_t(\theta)$. The model satisfies that

$$Ef_t(\theta) = 0 \ .$$

We will also refer to the vector f_t as moment conditions and θ is the K dimensional vector of parameters. The *identification* condition is that $Ef_t = 0$ for $\theta = \theta_0$ and otherwise not which, of course, implies $EM_T(\theta_0) = 0$. (Further you need to assume a compact parameter space or some equivalent assumption so there is not asymptote for, e.g., components of θ converging to infinity.)

Define

$$g_T \ = \ \frac{1}{T}M_T = \frac{1}{T}\sum_{t=1}^{T} f_t \ .$$

We will use the notation g_T or $g_T(\theta)$, but from now on the dependence of g_T on the underlying series, x_t, will be implicit. The GMM estimator will be the estimator that makes $g_T(\theta)$ as close to zero as possible. Notice that g_T is the empirical first moment of the series f_t which is why the estimator is called a moment estimator. Also note that the standard idea of moment estimation, which consists of equating as well as possible a series of moments. This would

be achieved by choosing $g'_T = [\overline{x}_T - Ex_t, \overline{x^2}_T - E\{x_t^2\}, ..., \overline{x^K}_T - E\{x_t^K\}]$.

We now define the **GMM-estimator** as

$$\hat{\theta} \;=\; \text{argmin}_\theta \; g'_T W_T g_T \;,$$

where W_T is a weighting matrix that (typically) depends on T such that there exist a positive definite matrix W_0, such that $W_T \to W_0$ $(a.s.)$. The latter condition allows us to let the weighting matrix be dependent on an initial consistent estimator, which is very important since the optimal GMM estimator will be a two step estimator, just as in the GLS-case above.

Let $Dg_T(\theta)$ be the $L \times K$ (number of moments $\times$ parameters) dimensional matrix of derivatives with typical element $Dg_{Tij} = \frac{\partial g_{Ti}}{\partial \theta_j}$. We will assume Dg_T has full rank. When the underlying data follows continuous distributions this will usually follow with probability 1 from the identification condition. (Below I will often just write Dg in order to simplify notation but, of course, all functions will be evaluated using the T available observations.)

Then the first order condition of the optimization becomes

$$Dg_T(\hat{\theta})' W_T \, g_T(\hat{\theta}) \;=\; 0 \;.$$

Solving non-linear optimization by the Newton algorithms

GAUSS and other programs use a Newton type algorithm to solve non-linear optimization problems. There are many variations of this but most variations involve approximations to how one finds derivatives and things like that. The computer will find the derivative of the criterion function numerically but you will have the option to let a subroutine calculate it if you have an analytical expression, this will often increase computational speeds significantly if the number of parameters is high.

Newton type algorithms work by starting from an initial value θ_0 and for a given value θ_{N-1} finding θ_N which minimize the linearized criterion function:

$$[g_T(\theta_{N-1}) + Dg(\theta_N - \theta_{N-1})]' W_T [g_T(\theta_{N-1}) + Dg(\theta_N - \theta_{N-1})]$$

The solution is (check this!)

$$\theta_N - \theta_{N-1} \;=\; -(Dg' W_T Dg)^{-1} Dg' W_T g_T(\theta_{N-1}) \;,$$

which is the NEWTON upgrade.

1.1 Asymptotic theory

We will assume that the series $(x_t', z_t')'$ is **ergodic**. A series x_t is ergodic if

$$\frac{1}{T}\sum_{t=1}^{T} h(x_t) \ \rightarrow Eh(x_t)$$

for all functions h(.) (for which the mean is well defined). Notice that the right hand side of the above equation is assumed to not be a function of t. It is, more or less, impossible to test if a series is ergodic. However, it is well known that an *integrated* time-series (e.g., a random walk) is not ergodic. Most macroeconomic series are integrated, or nearly integrated, time series, but most often the model can be rewritten in terms of stationary variables (typically growth rates).

For proof of consistency, see for example, Hansen (1982). The idea is simple enough. When T is large the function $g_T(\theta)$ is close to $Ef_t(\theta)$ and the minimum of g_T will therefore be close to the minimum of Ef_t, i.e., close to θ_0. In order to make these statements precise we need to be specific about what we mean by convergence of *functions* but I will leave this for more specialized econometrics courses.

We will also assume that the series $f_t(\theta)$ satisfies a central limit theorem,

i.e. that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t(\theta) \;\Rightarrow\; \mathrm{N}(0,\Omega)\ ,$$

where $\Omega = E[f_t f_t']$ if f_t is not autocorrelated, but in general

$$\Omega = \lim_{J\to\infty}\sum_{j=-J}^{J} E[f_t f_{t-j}']\ .$$

So intuitively, where we in the GLS model had T (or L in the IV case) normally distributed error terms, we here have L asymptotically normally distributed moment (or orthogonality) conditions.

Let $Dg = E\frac{\partial g_T}{\partial \theta}(\theta_0)$. (Notice that because g_T is an average of the f_t's, one could also write Df and the previous version of this note used both, but Dg seems more intuitive as it is the probability limit of Dg_T and just keep in mind that to cut down on clutter, we here use the convention that Dg without a T subscript is an expectation.) One can then show that for any convergent sequence of weighting matrices the GMM-estimator is consistent and asymptotically normal with

$$\sqrt{T}(\hat{\theta} - \theta\ ) \Rightarrow\ \mathrm{N}(0,\Sigma)\ ,$$

where

$$\Sigma \;=\; (Dg'W_0Dg)^{-1}Dg'W_0\Omega W_0Dg(Dg'W_0Dg)^{-1}\ .$$

Notice that this formula corresponds exactly to the one obtained in the linear case if you substitute X for Dg.

We can sketch the proof (see Hall's book for more detail): By the mean value theorem we can write

$$g_T(\hat{\theta}) = g_T(\theta_0) + Dg_T(\overline{\theta})(\hat{\theta} - \theta_0)$$

Now, pre-multiply this equation by $Dg_T(\hat{\theta})'W_T$ and, by the first-order condition above, the left-hand side is 0 and we get

$$0 \;=\; Dg_T(\hat{\theta})'W_Tg_T(\theta_0) + Dg_T(\hat{\theta})'W_TDg_T(\overline{\theta})\,(\hat{\theta} - \theta_0)\ ,$$

where $|\overline{\theta}-\theta_0| < |\hat{\theta}-\theta_0|$ (which implies that when $\hat{\theta}$ is consistent and converges to θ_0, so will $\overline{\theta}$). We get

$$\hat{\theta} - \theta_0 \;=\; -[Dg_T(\hat{\theta})'W_TDg_T(\overline{\theta})]^{-1}Dg_T(\hat{\theta})'W_Tg_T(\theta_0)\ .$$

This implies

$$\sqrt{T}(\hat{\theta} - \theta_0) \;=\; -[Dg_T(\hat{\theta})'W_TDg_T(\overline{\theta})]^{-1}Dg_T(\hat{\theta})'W_T\sqrt{T}g_T(\theta_0)$$

This has the form

$$A_T \sqrt{T} g_T(\theta_0) \ ,$$

where $\sqrt{T} g_T(\theta_0)$ converges in distribution to $\mathrm{N}(0, \Omega)$ and A_T converges in probability to $(Dg' W_0 Dg)^{-1} Dg' W_0$ which gives us the formula above. (Dg_T converges to $Dg (= Df)$ by ergodicity.) Again, you need to look at more specialized articles to make sure this is all kosher but, in general, the assumption that econometric theorists impose to prove the theorems are rarely of such a form that practitioners can verify them.

The reasoning behind the GLS estimator also carries over and the *optimal GMM-estimator* is the one where $W_T \to \Omega^{-1}$ in which case the asymptotic covariance of the GMM-estimator is

$$\Sigma_0 \ = \ (Dg' \Omega^{-1} Dg)^{-1} \ .$$

In order to obtain an estimate $\hat{\Sigma}_0$ you need an estimate $\hat{\Omega}$ and then you use

$$\hat{\Sigma}_0 = (Dg_T' \hat{\Omega}^{-1} Dg_T)^{-1} \ ,$$

where Dg_T, of course, is evaluated at $\hat{\theta}$. For a non-optimal weighting matrix,

you similarly estimate the variance as

$$\hat{\Sigma} \;=\; (Dg_T' W_0 Dg_T)^{-1} Dg_T' W_0 \hat{\Omega} W_0 Dg_T (Dg_t' W_0 Dg_T)^{-1} \;.$$

Notice that this is the optimal estimator for a **given set of instruments**. The problem of finding the best instruments is much harder and no satisfactory solution exists to that problem in general (although often for special cases, like the OLS model). I will comment on this (very important) issue below.

Also notice, that consistency is found without making assumption on the error terms and without specifying the model such that the error terms are independent. Sometimes authors will claim that this is a big strength of GMM, but if you error are not approximately normal you will often have problems and, in particular, if you have a lot of autocorrelation in your residuals you will not get very precise estimates. (The profession seems to go in circles as to whether it is consider a strength to not have to make

distributional assumptions ["OLS is the Best Linear Unbiased estimator" or the opposite "OLS is the ML estimator"]. Some people can get "religious" about this issue, but my more pragmatic attitude is that the important thing is to get error terms that are approximately uncorrelated.)

1.2 Hypothesis Testing in a GMM-framework

There exists equivalents of the standard Wald-, LM-, and ML-test in the case of GMM estimation. Note: This is only true in the case where the optimal weighting matrix has been applied. In a case where you apply a non-optimal weighting matrix then there is no equivalent of the ML-test available. (Ho, Perraudin, and Sørensen (1996) is an example of a paper that applies a non-optimal weighting matrix).

Consider a test for s nonlinear restrictions

$$R(\theta) \;=\; 0 \;,$$

where R is an $s \times 1$ vector of functions.

Let DR be $\frac{dR}{d\theta}$ (and we assume that DR is evaluated at the optimal GMM-

estimator in the unrestricted model), then the **Wald test** is

$$TR(\hat{\theta})'[DR\hat{\Sigma}DR']^{-1}R(\hat{\theta}) \ ,$$

or

$$TR(\hat{\theta})'[DR(Dg_T'WDg_T)^{-1}Dg_T'W\hat{\Omega}_TWDg_T(Dg_T'WDg_T)^{-1}DR']^{-1}R(\hat{\theta}) \ ,$$

where W is the weighting matrix and $Dg_T$is the derivative of g_T with respect to the parameters. Here Dg_T (and DR if this is dependent on the parameters) are evaluated at the *unrestricted* estimator of θ. We have

$$\hat{\Sigma} \ = \ (Dg_T'WDg_T)^{-1} \, Dg_T'W\hat{\Omega}_TWDg_T \, (Dg_T'WDg_T)^{-1} \ ,$$

In the formula for the Wald-test, $\hat{\Sigma}$ is our estimator of the variance of $\hat{\theta}$ and when we pre- and post-multiply this by DR we get an estimate of the asymptotic variance of $R(\hat{\theta})$ by the Delta rule.

The LM-test can be implemented in different ways. I strongly recommend you check with a trusted source (like the article in the handbook or Gallant (1987)). For example, there is a formula in Ogaki (1992) that I cannot quite get to agree with Gallant's formula and a much simpler looking formula in Davidson and MacKinnon, that I cannot see how they get. They

may be OK, but I recommend you be careful. The formula given here should agree with Gallant (1987). This version has the form

$$LM \;=\; Tg_T' W Dg_T (Dg_T' W Dg_T)^{-1} DR' (DR\hat{\Sigma} DR')^{-1} DR (Dg_T' W Dg_T)^{-1} Dg_T' W g_T$$

where Dg_T, g_T, and DR are evaluated at the *restricted* value $\hat{\theta}$.

One way to motivate this version of the LM test is notice that if the restriction $R(\theta)$ is true the $DR(\hat{\theta})d\theta$ (evaluated at the restricted estimator) should be approximately zero where $\hat{\theta}$ is evaluated at the constrained minimum. The idea (of this version of the LM-test) is that you choose $d\theta$ as the update in a NEWTON algorithm, i.e.,

$$\theta_N - \hat{\theta} \;=\; (Dg_T' W Dg_T)^{-1} Dg_T' W g_T(\hat{\theta}) \;,$$

If the model fits well or, rather, does not "want to" violate the restriction, the NEWTON step away from the restricted parameter value will be small or, at least, orthogonal to DR. Now you find the LM test-statistic by evaluating

$$[DR(\theta_N - \hat{\theta})]' V^{-1} DR(\theta_N - \hat{\theta}) \;,$$

where $V = DR\hat{\Sigma}DR'$ is the variance of $DR(\theta - \hat{\theta})$ (ignoring the small sample variance in DR), and

$$DR(\theta_N - \hat{\theta}) \;=\; DR(Dg_T' W Dg_T)^{-1} Dg_T' W_T g_T(\theta_{N-1}) \;,$$

using the expression for the Newton-step found above.

Finally the **LR-test** (of course it should strictly speaking be "LR-type test" for Likelihood-Ratio type) is

$$LR \;=\; 2 * T[J_T(\hat{\theta}^r) - J_T(\hat{\theta}^u)] \;,$$

where J_T is the objective function **(NB)** evaluated at the *optimal* weighting matrix and where the superscripts u and r of course indicates that the estimators were found in the unrestricted and the restricted models respectively.

The Wald-, LM-, and LR-test can all be shown to converge in distribution to a χ^2-distribution with s (number of restrictions) degrees of freedom in the case where the restrictions are true.

Hansen (1982) suggested the following **test for mis-specification**: Consider

$$J_T \;=\; T g_T(\hat{\theta})' \hat{\Omega}_T^{-1} g_T(\hat{\theta}) \;.$$

If the model is correctly specified this statistic is asymptotically χ^2 dis-

tributed with degrees of freedom equal to $L - K$—the degrees of freedom is number of moments minus number of parameters. A value that is far out in the tail indicates that the whole model is mis-specified. By the whole model, I do not mean that *all* parts of the model are mis-specified; but rather that *some* part of the model is mis-specified—it could be that it was just the instruments that were not pre-determined. This test is known as the **test for overidentifying restrictions** or sometimes as the "Hansen J-test".

Note that you cannot test unless you have more moment conditions than parameters (an "overidentified model"), in the case the model is exactly identified the J_T will be identically 0.

In Hansen and Singleton (1982) the model was rejected by the J-test, and my subjective impression is that from then on it became acceptable for a while to present an econometric estimation that rejected the model, as one that accepted the model. (This is the "scientific method" that Summers reject for macroeconomics. It seems that Summers won in that dimension, because at present it is basically impossible to publish an article that rejects the model.)

I often find the J-test useless. Models are never exactly true so the re-

sult of the J-test will usually be that it accepts the model (due to lack of power) if the number of observations is low, and rejects the model if the number of observations is high.

Simulated GMM

You may sometimes be in the situation where you cannot find an analytic expression for Ef_t. However, you might be able to *simulate* Ef_t. It is most easily explained by an example. Consider, for example, an $MA(2)$ process

$$x_t \;=\; u_t + b_1\, u_{t-1} + b_2 u_{t-2}\ , \quad (*)$$

where the error terms are $N(0, \sigma^2)$ distributed. The parameter vector here is $\theta' = \{b_1, b_2, \sigma^2\}$. For this model, I would actually use the Kalman-filter to evaluate the likelihood function for this particular model, but this is just an example, so imagine I couldn't find the likelihood function or the conditional likelihood function. Then I might simulate some moments for the y_t process. For example, I might use a random number generator to draw $N = 100,000$ observations $u_1, ..., u_N$ (actually, you need some initial values of u and y also, but that should be obvious to adjust for) and calculate

$y_1, ..., y_N$ using equation (*). I would set this up as a subroutine named, e.g., $SIM(\theta)$ in GAUSS [meaning that you call the subroutine SIM as a function of a given set of parameters]. Then, in the same subroutine I could calculate, say, $\tilde{m}_1(\theta) = \frac{1}{N}\Sigma_{i=1}^{N} y_i$, $\tilde{m}_2 = \frac{1}{N}\Sigma_{i=1}^{N} y_i y_{i-1}$ $\tilde{m}_3 = \frac{1}{N}\Sigma_{i=1}^{N} y_i y_{i-2}$, and $\tilde{m}_4 = \frac{1}{N}\Sigma_{i=1}^{N} y_i y_{i-3}$. (Note here that I changed the moments from the previous draft: in a dynamic model, you would get noisy estimates if you not use "dynamic moments.") Because N is a very large number, the m moments are very very close to the expected values and we can use them instead of those.

Assume now that x_t, $t = 1, ..., T$ is your actual data. You calculate the data moments for T observations the same way: $m_{T1}(\theta) = \frac{1}{T}\Sigma_{t=1}^{T} x_t$, $m_{T2} = \frac{1}{T}\Sigma_{t=1}^{T} x_t x_{t-1}$, etc. The parameter vector is $\theta = \{b_1, b_2, \sigma^2\}$. Now you would define $g_T' = \{\tilde{m}_1 - m_{T1}, \tilde{m}_2 - m_{T2}, \tilde{m}_3 - m_{T3}, \tilde{m}_4 - m_{T4}\}$. and you would minimize

$$\hat{\theta} \;=\; \text{argmin}_\theta \; g_T' W g_T \;,$$

for some weighting matrix W. This would give you a consistent estimate of θ. Note that this might be slow because for each step θ_N in the Newton algorithm, you need to call $SIM(\theta_N)$ in order to calculate the moments.

(NOTE here that you draw the random numbers once, not inside the SIM algorithm. If you do that, the simulated moments will wobble a little and the Newton algorithm might not settle down on a $\hat{\theta}$ estimate.) As a matter of fact, for this model this would not be a problem since a modern computer can do this very quickly. In principle, *any* model that can be simulated (which more or less is the universe of models can be put into the SIM routine and some moments returned). In practice, you would have trouble with a large General Equilibrium (GE) model—GE models typically would need to be simulated which means that you would add a layer of non-linear simulations for each θ_N...but as computers get faster you might be able to do it for a small GE model if you program cleverly. (Since there would be billions of calculations they better be streamlined.)

You need to choose N much larger than T—otherwise you need to take the extra variance that comes from simulating the moments into account when calculating std. errors. If your moments come from simulating a complex model, you may be forced to have N small.

Choice of moments. What matter much more for efficiency than the choice of weighting matrix is the choice of moments. In the case, as in the Hansen-Singleton model, where "choice of moments" means "choice of instruments" theory gives little guidance. You can show that as T gets larger you should use more instruments, but in practice you have one T and you have to use common sense (use instruments that are not to correlated, don't use too many, ...). In the case, such as the MA(2) example, where you actually choose moments, you can more or less guess which moments will be good. For example, the ones I chose above, were pretty bad. An MA(2) model is characterized by non-zero first and second autocorrelations and higher order correlations being zero. So good moments would be the empirical variance, first, second, third, and maybe fourth order autocorrelations, rather than the higher moments I chose above.

It is possible to be quite systematic about this. Gallant and Tauchen suggested a method called **Efficient Method of Moments** (EMM) that can be used if you have a model with a likelihood function that you cannot write down such as a stochastic volatility model but you have a model that captures similar features of the data such as a GARCH model, you can ac-

tually estimate the GARCH model even if it is misspecified and then use the first derivatives of the likelihood function as the moment conditions. More precisely, if $s(\theta, x)$ is the score function (the derivative of the likelihood function) as a function of the data, you use simulated method of moments to match this to the model, but drawing a long series of observations y_i and calculate $s(\theta, y)$ and then your moment conditions are $g_T = s(\theta, x) - s(\theta, y)$. For the particular models that I mentioned this turns out to actually work very well—see the very comprehensive Monte Carlo papers by Andersen and Sorensen (1996) and Andersen, Chung, and Sorensen (1999). I also have some hand-written notes on EMM that you can have if you are interested.

Literature: Gallant (1987)

Newey and McFadden: Large Sample Estimation and Hypothesis Testing. In Handbook of Econometrics IV, eds. Engle and McFadden, North-Holland, 1994.