Bent E. Sørensen
January 23, 2007

1 Teaching notes on GMM 1.

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). For introductions to GMM, Davidson and MacKinnon (1993) have comprehensive chapter on GMM and I recommend that you read the chapter on GMM in the Hamilton (1994) textbook. This is a good supplement to the teaching notes. For more comprehensive coverage see the recent textbook 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, since that will contain most of the ideas and intuition for the general GMM estimation.

1.1 Linear IV estimation

Consider the following simple model

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

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 studies.
In vector form the equation (1) can be written

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

in the usual fashion. If x_t and e_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 e_t — so if x_t is actually uncorrelated with e_t you will use x_t itself as instruments - in this way all the simple estimators that you know, like OLS, are special cases of IV- (and GMM-) estimation. If Z denotes the $T \times L$ $(K \geq L)$ vector of the z-observations then we get by pre-multiplying (2) by Z that

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

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

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

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

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

Now the standard OLS estimator of θ is

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

which is consistent and unbiased with variance

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

For simplicity let us now consider drop the tilde's, and just remember that the system (of the form (2)) often will have been obtained via the use of instrumental variables. (Most of the GMM-literature uses very sparse notation, which is nice when you are familiar with it, but makes it hard to get started on).
If U does not have a variance matrix that is proportional to the identity matrix the OLS estimator is not efficient. Remember that the OLS estimator is chosen to minimize the criterion function

$$U'U \ = \ (Y - X\theta)'(Y - X\theta) \ .$$

To obtain a more **efficient** estimator than the OLS 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

$$U'WU \;=\; (Y - X\theta)'W(Y - X\theta) \;,$$

or (in the typical compact notation)

$$\hat{\theta} \;=\; argmin_{\theta} U'WU \;.$$

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

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

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

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

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

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

We recognize this as the variance 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.

For practical purposes one would usually have to do a 2-step estimation. First perform a preliminary estimation by OLS (for example), then estimate Ω (from the residuals), and perform a second step using this estimate of Ω to perform "feasible GLS". 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 $var(E) = I$, then the variance of $Z'E$ is $Z'Z$. 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 $Z(Z'Z)^{-1}Z'X$. 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 estimation

$$(4) \quad Y = h(X,\theta) \;+\; U \;,\;,$$

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

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

In GMM jargon the model would usually be formulated as

$$U \;=\; Y - h(X,\theta) \;,$$

or more often as

$$(**) \quad U \;=\; f(\tilde{X},\theta) \;,$$

(where $\tilde{X} = Y, X$ and $f(\tilde{X},\theta) = Y - X\theta$. The later—very compact—notation is the one that is commonly used in the GMM literature and we will follow it here. We again drop the tilde and denote all the variables by X. 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**:

$$EU \;=\; 0 \;,$$

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

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

Here you should think of U as being a theoretical model. It is not quite explicit here whether we think of U as equations that have been pre-multiplied by instrument vectors or

not. But in the usual formulation of GMM the dimension L of U is fixed, so e.g. in the OLS model where the dimension of $E = \{e_1, ..., e_T\}'$ depends on T, you would think of the orthogonality conditions as being $U = X'Y - X'X\theta$. 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 $f(X, \theta)$ that satisfies $Ef(X, \theta) = 0$, the GMM estimator attempts to minimize a quadratic form in f, namely $f'Wf$. 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 $(Y - X\theta)$]. Assume that Z is just columns of ones. Then a relation like $f(X, \theta) = Z'g(X, \theta)$ is just $g_T(X, \theta) = \frac{1}{T}\Sigma_{t=1}^{T} g_t(X, \theta)$. In other words the orthogonality condition is the first empirical moment of the g_t vector. In the case of instruments z_t the orthogonality condition is really $g_T(X, \theta) = \frac{1}{T}\Sigma z_t g_t(X, \theta)$. If the number of orthogonality conditions is the same as the number of parameters you can solve for the θ vector which makes $g_T = 0$ – 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 rep-

resentative 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

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

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

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

C_t	:	Consumption
I_t	:	Investment in (one-period) asset
W_t	:	Other Income
r_t	:	Rate of Return
β	:	Discount Factor
γ	:	Parameter of Utility Function

If you knew how C_t and I_t was determined this model could be used to find r_t (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

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

The model can 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 can be used to form the "orthogonality condition"

$$U(C_t, \theta) \;=\; \beta C_{t+1}^{\gamma-1} r_{t+1} \;-\; C_t^{\gamma-1} \;,$$

where $E_t U = 0$ (why?), where we now interpret E_t as the "objective" or "true" conditional expectation. Note that $E_t U = 0$ implies that $EU = 0$ 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 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. Ω_t will typically be a set of previous observations of economic variables $\{z_{1t}, z_{1t-1}, ...; z_{2t}, z_{2t-1}, ...; z_{Kt}, z_{Kt-1}, ...\}$. (Including C_t, and I_t among the z's. Then any variable in Ω_t will be a valid instrument in the sense that

$$E[z_t U(C_t, \theta)] \;=\; 0$$

for any z_t in Ω_t. Notice that z_t here denotes any valid instrument at time t, for example z_t could be z_{1t-3} - this convention 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 not are very efficient if no good instruments are available (there is active research in this area at the present, see paper with the words "weak instruments"); but for now you may want to compare the Hansen-Singleton (1982) approach to the article "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns", JPE, 93, p 249-265. This is really the same model, but with enough assumptions imposed that the model can be estimated by Maximum Likelihood.