

LECTURE 3A: OTHER NONLINEAR ESTIMATION (GMM AND EM)

CHRIS CONLON

NYU STERN

MARCH 7, 2019

ESTIMATING FINITE MIXTURES

- In practice estimating finite mixture models can be tricky.
- A simple example is the mixture of normals (incomplete data likelihood)

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^N \sum_{k=1}^K \pi_k f(x_i | \mu_k, \sigma_k)$$

- We need to find both mixture weights $\pi_k = Pr(z_k)$ and the components (μ_k, σ_k) the weights define a valid probability measure $\sum_k \pi_k = 1$.
- Easy problem is **label switching**. Usually it helps to order the components by say decreasing $\pi_1 > \pi_2 > \dots$ or $\mu_1 > \mu_2 > \dots$.
- The real problem is that which component you belong to is unobserved. We can add an extra indicator variable $z_{ik} \in \{0, 1\}$.
- We don't care about z_{ik} per-se so they are **nuisance parameters**.

- We can write the complete data log-likelihood (as if we observed z_{ik}):

$$l(x_1, \dots, x_n | \theta) = \sum_{i=1}^N \log \left(\sum_{k=1}^K I[z_i = k] \pi_k f(x_i | \mu_k, \sigma_k) \right)$$

- We can instead maximize the expected log-likelihood where we take the expectation $E_{z|\theta}$

$$\alpha_{ik}(\theta) = \Pr(z_{ik} = 1 | x_i, \theta) = \frac{f_k(x_i, \mu_k, \sigma_k) \pi_k}{\sum_{m=1}^K f_m(x_i, \mu_m, \sigma_m) \pi_m}$$

- Now we have a probability $\hat{\alpha}_{ik}$ that gives us the probability that i came from component k . We also compute $\hat{\pi}_k = \frac{1}{N} \sum_{i=1}^N \alpha_{ik}$

- Treat the $\hat{\alpha}_k(\theta^{(q)})$ as data and maximize to find μ_k, σ_k for each k

$$\hat{\theta}^{(q+1)} = \arg \max_{\theta} \sum_{i=1}^N \log \left(\sum_{k=1}^K \hat{\alpha}_k(\theta^{(q)}) f(x_i | z_{ik}, \theta) \right)$$

- We iterate between updating $\hat{\alpha}_k(\theta^{(q)})$ (E-step) and $\hat{\theta}^{(q+1)}$ (M-step)
- For the mixture of normals we can compute the M-step very easily:

$$\begin{aligned} \mu_k^{(q+1)} &= \frac{1}{N} \sum_{i=1}^N \hat{\alpha}_k(\theta^{(q)}) x_i \\ \sigma_k^{(q+1)} &= \frac{1}{N} \sum_{i=1}^N \hat{\alpha}_k(\theta^{(q)}) (x_i - \bar{x})^2 \end{aligned}$$

- EM algorithm has the advantage that it avoids complicated integrals in computing the expected log-likelihood over the missing data.
- For a large set of families it is proven to converge to the MLE
- That convergence is **monotonic** and **linear**. (Newton's method is quadratic)
- This means it can be slow, but sometimes $\nabla_{\theta} f(\cdot)$ is really complicated.

Generalized Method of Moments is a powerful tool to estimate econometric models under weaker assumptions than MLE:

- We no longer need to know $f(y, x|\theta)$.
- Instead we just need that $E[g(y_i, x_i; \theta)] = 0$ (**Moment Condition**)
- Like MLE this is another **M-Estimator** or **Extremum Estimator**
- We choose $\hat{\theta}_{GMM}$ to minimize some objective function that is an expectation over our entire dataset.

- some data w_i where $i = 1, \dots, N$
- Our data w_i might contain all kinds of things such as dependent variables y_i , regressors x_i and excluded instruments z_i .
- Our economic model provides the following restriction on our data:

$$E[g(w_i, \theta_0)] = 0$$

- At the true parameter value $\theta_0 \in \mathbb{R}^k$ our moment conditions $g(w_i, \theta)$ are on average equal to zero.
 - ▶ What does “on average” mean?
 - ▶ In theory, we are making an asymptotic statement about what happens as $N \rightarrow \infty$. This is what we mean when we write $E[\cdot]$.

GMM SETUP

In practice, it is helpful to consider the sample analogue:

- We write as $g_N(\theta) \in \mathbb{R}^q$, where $g_N(\theta)$ is a q -dimensional vector of moment conditions:

$$E[g(w_i, \theta)] \approx \frac{1}{N} \sum_{i=1}^N g(w_i, \theta) \equiv g_N(\theta)$$

- We define the Jacobian: $D(\theta) \equiv E[\frac{\partial g(w_i, \theta)}{\partial \theta}]$, which is a $q \times k$ matrix.
- Evaluated at the optimum, $\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \xrightarrow{d} N(0, S)$ where $S = E[g(w_i, \theta_0)g(w_i, \theta_0)']$ is a $q \times q$ matrix.
- In other words, the moment conditions which are 0 in expectation at θ_0 are normally distributed with some covariance S .
- Later, we will refer to a weighting matrix W_N which is a $q \times q$ positive semi-definite matrix. It tells us how much to penalize the violations of one moment condition relative to another (in quadratic distance).

GMM: EXAMPLES

Here is the GMM estimator:

$$\hat{\theta} = \arg \min_{\theta} Q_N(\theta) \quad Q_N(\theta) = g_N(\theta)' W_N g_N(\theta)$$

It is easy to see some very simple examples:

OLS Here $y_i = x_i\beta + \epsilon_i$. Exogeneity implies that $E[x_i'\epsilon_i] = 0$. We can write this in terms of just observables and parameters as $E[x_i'(y_i - x_i\beta)] = 0$ so that $g(y_i, x_i, \beta) = x_i'(y_i - x_i\beta)$.

IV Again $y_i = x_i\beta + \epsilon_i$. Now, endogeneity implies that $E[x_i'\epsilon_i] \neq 0$. However there are some instruments z_i which may be partly contained in x_i and partly excluded from y_i , so that $E[z_i'\epsilon_i] = 0$. $E[z_i'(y_i - x_i\beta)] = 0$ so that $g(y_i, x_i, z_i, \beta) = z_i'(y_i - x_i\beta)$.

Maximum Likelihood $g(w_i, \theta) = \frac{\partial \log f(w_i, \theta)}{\partial \theta}$ where $f(w_i, \theta)$ is the density function so that $\log f(w_i, \theta)$ is the contribution of observation i to the log-likelihood. Here we set the expected (average) derivative of the log-likelihood (score) function to zero.

A FAMOUS EXAMPLE: EULER EQUATION

Assume we have a CRRA utility function $u(c) = \frac{c^{1-\gamma}-1}{1-\gamma}$ and an agent who maximizes the expected discounted value of their stream of consumption. This leads to an **Euler Equation**:

$$E \left[\beta \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} R_{t+1} - 1 | \Omega_t \right] = 0$$

where Ω_t is the **Information Set** (sigma algebra) of everything known to the agent up until time t (include full histories). We can write a moment restriction of the form for any measurable $z_t \in \Omega_t$.

$$E \left[z_t \left(\beta \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} R_{t+1} - 1 \right) \right] = 0$$

In the original work by Hansen (1982) on GMM, this $g(c_t, c_{t+1}, R_{t+1}, \beta, \gamma)$ was used to estimate (β, γ) .

GMM: TECHNICAL CONDITIONS

Here is the GMM estimator:

$$\hat{\theta} = \arg \min_{\theta} Q_N(\theta) \quad Q_N(\theta) = g_N(\theta)' W_N g_N(\theta)$$

These are a set of sufficient conditions to establish consistency and asymptotic normality of the GMM estimator. These conditions are stronger than necessary, but they establish the requisite LLN and CLT.

1. $\theta \in \Theta$ is compact.
2. $W_N \xrightarrow{p} W$.
3. $g_N(\theta) \xrightarrow{p} E[g(z_i, \theta)]$ (uniformly)
4. $E[g(z_i, \theta)]$ is continuous.
5. We need that $E[g(z_i, \theta_0)] = 0$ and $W_N E[g(z_i, \theta)] \neq 0$ for $\theta \neq \theta_0$ (global identification condition).
6. $g_N(\theta)$ is twice continuously differentiable about θ_0 .
7. θ_0 is not on the boundary of Θ .
8. $D(\theta_0) W D(\theta_0)'$ is invertible (non-singular).
9. $g(z_i, \theta)$ has at least two moments finite and finite derivatives at all $\theta \in \Theta$.

The first five conditions give us consistency $\hat{\theta} \xrightarrow{p} \theta_0$ as $N \rightarrow \infty$. All nine conditions give us asymptotic normality.

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V_{\theta})$$

SPECIAL CASE: LINEAR MODEL

The global identification condition is difficult to understand, for the linear model we can replace it with a (local) condition on Jacobian of the moment conditions.

- Recall the Jacobian: $D \equiv \frac{\partial g(w_i, \theta)}{\partial \theta}$, which is a $q \times k$ matrix.
- For OLS this reduces to $(X'X)$
- We call the problem:
 - Under-identified** if $\text{rank}(D) < k$
 - Just-identified** if $\text{rank}(D) = k$
 - Over-identified** if $\text{rank}(D) > k$.
- In the under-identified case, there may be many such $\hat{\theta}$ where $g(w_i, \hat{\theta}) = 0$.

We are primarily interested in the **over-identified case**

$$\hat{\theta} = \arg \min_{\theta} Q_N(\theta) \quad Q_N(\theta) = g_N(\theta)' W_N g_N(\theta)$$

- There is no $\hat{\theta}$ which satisfies the moment conditions $g_N(\hat{\theta}) \neq 0$.
- Instead, we search for $\hat{\theta}$ which minimizes the violations of the moment conditions.
- We write this as a quadratic form for some positive definite matrix W_N which is $q \times q$.

For the linear IV problem this becomes:

$$\begin{aligned}g_N(\theta)'W_Ng_N(\theta) &= \frac{1}{N^2} \cdot (\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\beta))'W_N(\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\beta)) \\&= \frac{1}{N^2} \cdot [Y'ZW_NZ'Y - 2\beta X'ZW_NZ'Y + \beta'X'ZW_NZ'X\beta]\end{aligned}$$

We can ignore the $\frac{1}{N^2}$ and take the first-order condition:

$$\begin{aligned}2X'ZW_NZ'Y &= 2X'ZW_NZ'X\beta \\ \hat{\beta}_{GMM} &= (X'ZW_NZ'X)^{-1}X'ZW_NZ'Y\end{aligned}$$

Suppose that we do not have any excluded instruments so that $Z = X$ (and thus $q = k$). Also suppose that $W_N = \mathbf{I}_q$ (the identity matrix). Then we can see that:

$$\begin{aligned}\hat{\beta}_{GMM} &= (X'X\mathbf{I}_qX'X)^{-1}X'X\mathbf{I}_qX'Y \\ &= (X'XX'X)^{-1}X'XX'Y \\ &= (X'X)^{-1}(X'X)^{-1}(X'X)X'Y = (X'X)^{-1}X'Y = \hat{\beta}_{OLS}\end{aligned}$$

- In other words, OLS is a special case of the GMM estimator.
- Also, the identification condition $D = \frac{\partial g(w_i, \theta)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^N z_i' x_i = \frac{1}{N} \sum_{i=1}^N x_i' x_i$ becomes that $\text{rank}(X'X) = k$ the well-known OLS rank condition

2SLS ESTIMATOR

Suppose that we do have excluded instruments so that:

- $\dim(Z) = q > \dim(X) = k$
- $W_N = (Z'Z)^{-1}$

It immediately follows that:

$$\hat{\beta}_{GMM} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y = \hat{\beta}_{2SLS}$$

If $\dim(Z) = q = \dim(X) = k$ then $(X'Z)$ is square (and invertible). This expression further simplifies:

$$\begin{aligned}(X'Z(Z'Z)^{-1}Z'X)^{-1} &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1} \\ \rightarrow \hat{\beta}_{GMM} &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'Y = (Z'X)^{-1}Z'Y = \hat{\beta}_{IV}\end{aligned}$$

CHOICE OF WEIGHTING MATRIX

How do we choose the weighting matrix W_N . Already seen two options:

1. The identity matrix \mathbf{I}_q equally penalizes violations of all q moments;
2. TSLS weighting matrix $(Z'Z)^{-1}$ which can be thought about as the inverse of the covariance of the instruments.

We are interested in **efficient GMM** which is the GMM estimator with the lowest variance.

In order to find the W_N which minimizes the variance of $\hat{\theta}_{GMM}$ we recall the asymptotic variance (**sandwich form**) of the GMM estimator:

$$V_{\theta} = \underbrace{(DWD')^{-1}}_{\text{bread}} \underbrace{(DWSW'D')}_{\text{filling}} \underbrace{(DWD')^{-1}}_{\text{bread}}$$

It turns out that the best choice of $W_N = S^{-1}$ (which sets *filling* = *bread*). This is easy to see, because W_N is positive semi-definite.

$$(DS^{-1}D')^{-1}(DS^{-1}SS^{-1}'D')(DS^{-1}D')^{-1} = (DS^{-1}D')^{-1}(DS^{-1}D')(DS^{-1}D')^{-1} = (DS^{-1}D')^{-1}$$

What are we looking for, in ideal world:

- S to be small (we want the sampling variation/noise of our moments to be as small as possible)
- D (the Jacobian of the moments) to be large. This means that small violations in moment conditions lead to large changes in the objective function.
- The problem is well identified when the objective function is steep around θ_0 .
- When the problem becomes flat, it becomes hard to distinguish one θ in favor of another.

The problem is that $S = E[g(w_i, \theta_0)g(w_i, \theta_0)']$ is not something that we readily observe from our data. In fact, the asymptotic covariance evaluated at θ_0 is **infeasible**. The best we can hope for is to use some sample analogue $W_N = \hat{S}^{-1}$ in its place. One way to compute that is the covariance of the moments estimated at some $\hat{\theta}$ for an initial guess of W :

$$\hat{W} = \hat{S}^{-1} = \left(\frac{1}{N} \sum_{i=1}^N (g(w_i, \hat{\theta}) - g_N(\hat{\theta})) (g(w_i, \hat{\theta}) - g_N(\hat{\theta}))' \right)^{-1}$$

Because $E[g(w_i, \theta_0)] = 0$ at θ_0 there is a tendency to use $(\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) g(w_i, \hat{\theta})')^{-1}$ (without de-meaning the moments). In theory this would work fine, but in practice **it is nearly always a bad idea**.

The overall procedure works as follows:

1. Pick some initial weighting matrix W_0 : often \mathbf{I}_q or $(Z'Z)^{-1}$.
2. Solve $\hat{\theta} = \arg \min_{\theta} g_N(\theta)' W_0 g_N(\theta)$.
3. Update $\hat{W} = \left(\frac{1}{N} \sum_{i=1}^N (g(w_i, \hat{\theta}) - g_N(\hat{\theta})) (g(w_i, \hat{\theta}) - g_N(\hat{\theta}))' \right)^{-1}$
4. Solve $\hat{\theta}_{GMM} = \arg \min_{\theta} g_N(\theta)' \hat{W} g_N(\theta)$.
5. Compute $D(\hat{\theta}_{GMM})$ and $S(\hat{\theta}_{GMM})$ and compute standard errors.

GMM: EXAMPLE

For the linear IV estimator when i is independent then $g(w_i, \theta) = z_i \epsilon_i$ and $E[z_i \epsilon_i] = 0$

$$\hat{S} = \frac{1}{N} \sum_{i=1}^N z_i z_i' \epsilon_i^2$$

- With homoskedastic variance $E[\epsilon_i^2 | z_i] = \sigma^2$ and the covariance of the moments becomes $\frac{\sigma^2}{N} \sum_{i=1}^N z_i z_i'$.
- Because scaling weighting matrix by a constant has no effect on the maximum this is equivalent to the 2SLS weight matrix: $\sum_{i=1}^N z_i z_i'$ or $Z'Z$.
- Thus 2SLS is only the efficient estimator when **homoskedasticity** is a reasonable assumption.
- If all regressors are exogenous then $X = Z$ and we are left with the GMM formula coincides with the covariance for heteroskedasticity robust standard errors.
- Similarly, when appropriate we can consider extensions such as **clustered standard errors** which are robust to weaker forms of independence.
- As a practical matter, we should always use the *sandwich* form when calculating the GMM standard errors, rather than the simpler *bread* version which is only correct at θ_0 under asymptotic optimality conditions.

In a famous paper, Chamberlain (1987) showed that GMM obtained the **semi-parametric efficiency bound** asymptotically.

- as our sample gets large and without making additional parametric assumptions the efficient GMM provides the most efficient estimator.
- We get close to MLE benefits without distributional assumptions!

1ST ESTIMATES LOOK OK.. 2ND STAGE ESTIMATES LOOK LIKE GARBAGE

This means there is a **problem with your weighting matrix!**

- Make sure you are subtracting off the average $g_N(\theta)$ when computing the covariance.
- Another problem appears is when you take the inverse. There is a statistic known as the **condition number** which measures the ratio of the minimum and maximum eigenvalue of a matrix.
 - ▶ When these eigenvalues are close together then small errors in $A + \epsilon$ lead to small errors in $(A + \epsilon)^{-1}$.
 - ▶ Software sometimes reports either the condition number, or its inverse.
 - ▶ In an ideal world this would be ≈ 1 .
 - ▶ If the condition number is $10^{\pm 13}$ it approaches the numerical precision of your computer.
 - ▶ Even tiny (sampling) errors can lead to nearly infinite weighting matrices.
- Usually this happens if the gradient of $Q_n(\theta)$ is not close to zero in a particular dimension or the average Jacobian $D(\theta) = \frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \theta)}{\partial \theta}$ is not close to zero in some dimension. You are not at the optimum of $Q_n(\theta)$!

POINT ESTIMATES GOOD, SE'S NOT SO MUCH...

- Assuming it is not one of the problems described above, you have probably dropped a $\frac{1}{N}$ or $\frac{1}{\sqrt{N}}$ somewhere.
- You can ignore the N 's when obtaining point estimates because scaling $Q_n(\theta)$ or W_n does not affect the location of $\hat{\theta}$, but you need to be more careful in computing V_θ .

WHY NOT 3 OR 4 STEP GMM?

Can I get even smaller standard errors?

- Asymptotically the answer is no. Even with a sub-optimal choice of weighting matrix, your first stage $\hat{\theta} \xrightarrow{P} \theta_0$ as $N \rightarrow \infty$.
- This means that in the limit, you always recover the efficient choice of \hat{W} .
- In finite sample, anything can happen, but numerous studies have found little to no improvement from considering a K -step GMM estimator.

WHAT ABOUT INFINITE STEP GMM?

There isn't such a thing. But there is the Continuously Updating Estimator (CUE) of Hansen Heaton and Yaron (1996). In this case we let:

$$W(\theta) = \left(\frac{1}{N} \sum_{i=1}^N (g(w_i, \theta) - g_N(\theta))(g(w_i, \theta) - g_N(\theta))' \right)$$
$$Q_n(\theta)^{CUE} = g_N(\theta)' \left(\frac{1}{N} \sum_{i=1}^N (g(w_i, \theta) - g_N(\theta))(g(w_i, \theta) - g_N(\theta))' \right) g_N(\theta)$$

- Because the weighting matrix now changes with θ , even for linear models this becomes very difficult to estimate.
- The original GMM problem was a quadratic optimization problem.
- The CUE problem is **no longer a convex optimization problem**
- This means that even numerical Quasi-Newton approaches are **no longer guaranteed to work**

In practice, CUE is not very popular for this reason.

MORE ON CUE

CUE has some additional advantages. We see this by examining the GMM FOC.

$$\hat{D}(\theta)W_n\hat{g}_N(\hat{\theta}) = 0$$

If we take an **Edgeworth expansion** we can see how the **asymptotic bias** term looks:

$$\frac{1}{N^2} \sum_{i=1}^N E[g(w_i, \theta)W_n g(w_i, \theta)]$$

- As long as the variance is finite we can see that this bias disappears as $N \rightarrow \infty$.
- it also tends to grow linearly with q the number of moment restrictions. This is often referred to as the **too many moments** or **many instruments** problem.
- When we estimate $\hat{g}(\hat{\theta})$ and $\hat{D}(\hat{\theta})$ we construct them so that they are mechanically correlated with one another.
- It turns out the CUE estimator fixes this in exactly the right way.
- How is beyond the scope of this note (and course). If you are really interested you should look at Newey and Smith (2004).

EXAMPLE: GRAVITY EQUATION

- An important set of models in international trade talk about **Gravity Equations**
- They are called gravity because trade declines with distance (or distance²).

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij}$$

Take Logs

$$\ln T_{ij} = \ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j + \alpha_3 \ln D_{ij} + \ln \eta_{ij}$$

- T_{ij} (Exports from i to j)
- (Y_i, Y_j) GDP of each country
- D_{ij} distance between two countries

EXAMPLE: GRAVITY EQUATION

If the moment condition holds then everything is good:

$$E[\ln(\eta_{ij})|Y_i, Y_j D_{ij}] = 0$$

Some problems

- Lots of Zeros in T_{ij} (so we can't take logs).
- If $(Y_i, Y_j D_{ij}, \eta_{ij})$ has heteroskedasticity then moment condition is violated.
- Why? Expectation is **linear operator** but $\log(\cdot)$ not so much.

Rearrange things so that:

$$T_{ij} = \exp(\beta_0 + \alpha_1 \log(Y_i) + \alpha_2 \log(Y_j) + \alpha_3 \log(D_{ij})) \eta_{ij}$$

$$T_{ij} = \exp(x_i \beta) \eta_{ij}$$

This gives us our moment condition:

$$E[T_{ij} - \exp(x_i \beta) | x_i] = 0$$

This works as long as we are okay with **proportional variance**