

These are some of my supplemental notes and exercises for EC484, the core course of the LSE EME. Basically, these are the notes I wish I had had at the beginning of last year. I don't think they give anything away, nor will they alleviate substantially the suffering that is the LSE EME, they just comment and expand on the class notes given.

For being a set of notes that complain so much about mistakes in the book, I'm sure they're full of mistakes. I apologize for that. Please let me know if you discover any errors, find something confusing, or just want to chat about econometrics.

- Ryan Giordano

1 Michelmas Term

1.1 0.6 The Partitioned Matrix Inverse

If you have a hard time remembering these partitioned matrix inverse formulas, there is an easy way to re-derive them using the same method you learned in high school for inverting scalar matrices by hand. Recall first that if the partitions are conformal, then the rules for multiplying partitioned matrices look just like the rules for multiplying scalar matrices:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{pmatrix}$$

(Here and below I'm not going to be pedantic – please assume everything I multiply is conformal, and everything I invert is invertible. This may get a bit confusing because I use "0" to refer to matrices containing all zeros but of different sizes, but the size should be clear from context, and I don't want this to take all day to do.)

Similar to the scalar case, you can multiply by clever matrices to add multiples of rows or columns to other rows and columns:

$$\begin{pmatrix} A & 0 \\ D & B \end{pmatrix} \begin{pmatrix} I_A & C \\ 0 & I_B \end{pmatrix} = \begin{pmatrix} A \times I_A + 0 \times 0 & A \times C + 0 \times I_B \\ D \times I_A + B \times 0 & D \times C + B \times I_B \end{pmatrix} = \begin{pmatrix} A & AC \\ D & B + CD \end{pmatrix}$$

Which has the effect of adding a right multiple of the first column to the second column. Similarly, you can add left multiples of one row to another, left multiple an entire row or right multiply a column by a matrix.

Given this, and the identity

$$A^{-1}A = I_A$$

if one can find a sequence of matrices Q_1, Q_2 , etc. such that

$$Q_1Q_2Q_3\dots Q_nA = I$$

then it follows that

$$Q_1Q_2Q_3\dots Q_nI = A^{-1}$$

If these Q_i matrices are of the form above, that is, which have the effect of adding multiples of rows to other rows (since we are left multiplying here), then this means that applying those same operations to the identity matrix will give you the inverse of A . (Here, it gives you the left inverse, but remember that the left inverse of a matrix is also the right inverse.) This is hopefully familiar from your introductory linear algebra – my point is just that you can do the same basic thing with partitioned matrices.

To do a simple example, let's invert

$$\begin{pmatrix} A & B \\ B' & 0 \end{pmatrix}$$

where B is not necessarily square. The steps are

1. Add $-B'A^{-1}$ times row one to row two
2. Add $B(B'A^{-1}B)^{-1}$ times row two to row one
3. Multiply the first row by A^{-1}

4. Multiply the second row by $-(B'A^{-1}B)^{-1}$

This might look a little mysterious, but if you work it out yourself you'll see that each step is pretty obvious, and that by performing these steps on the identity matrix, you get the exact same answer as in the book for the case $A_{22} = 0$. This method is easy to memorize, however, and quite general. Furthermore, in an exam, you usually don't need the full inverse partitioned form in all its glory – usually, you have some special case, in which case the above steps might be easy to do quickly.

1.2 Asymptotic Unbiasedness

It's a strange idea that a sequence of random variables can have an unbiased limiting distribution (in that the mean of a random variable with the limiting distribution is the true mean), and yet the sequence can still be asymptotically biased. But it can. Try to think of such a sequence.

1.3 Uniform Integrability

Think of a sequence that meets the necessary condition but is not U.I.

1.4 Section 4, Theorem 12

I haven't had enough serious probability theory to know the answer to this one, but it sure seems to me that you can't just go interchanging the order of finite sums of infinite sums of random variables and get the same random variable without some extra conditions attached. (You can't generally do it in the case of ordinary real numbers unless the sum is absolutely summable, for instance.) I never did get a very satisfactory answer to this question, and if you do, please email me.

It's worth thinking carefully about why you have to split the infinite sum into two pieces in the proof of Theorem 12.

1.5 Section 5, Almost Sure Convergence

When I first read this section, I didn't get what the point was. Let me call condition (1) in theorem 13 "Bounded Probability". Then the point of this section is to show that

- Bounded Probability \Rightarrow The almost sure law of large numbers (not proven in the course notes)
- $1 + \eta$ absolute mean convergence \Rightarrow Bounded Probability (proven by construction in the section starting just before Example 14)
- Bounded Probability \Rightarrow Uniform Integrability (proven directly in the first part after theorem 13)

Note that if X_i are iid, then the variable $Y = |X_1|$ automatically satisfies the bounded probability criterion with $C = 1$, so iid sequences with $E|X_i| < \infty$ converge with the strong LLN, which you probably already knew.

1.6 Section 7, Example 16

In my humble opinion, the reasoning does not look right here. In the section where he uses $tr(S_i) = O(i^\alpha)$ inside a big-O operator, he is implicitly saying something like

$$\begin{aligned}x_n &= O_p(g(y_n)), \\y_n &= O_p(z_n) \Rightarrow \\x_n &= O_p(g(z_n))\end{aligned}$$

which is not true generally. Try to find a counterexample.

1.7 Central Limit Theorems

You might not find my comments here very useful, but my main point is that this is not as difficult to do as the notes make it look. For me, the central confusing point was the notation. It may be easier to think in terms of the following steps, grasp the intuitive argument, and then go back afterwards to see which parts are which.

One of the main formal difficulties of doing CLTs in this course are that, in order to use any of the theorems, you need to know the exact variance of the thing you're trying to do a CLT on. In general, estimators have nasty complicated expressions with inverses in them and you cannot take expectations of the entire expression. The first step is separating out the stochastic and non-stochastic pieces (you have lots of tools – add and subtract or multiply and divide by limits, and make sure you do a CLT only on the smallest part you

absolutely have to, and then use the CMT for the rest). Then you have to normalize by the true variance so that the resulting variance is one. In the notes, this part is easy, and as I recall, in the homework is harder.

The second difficulty is to show that you have the appropriate rates of convergence. Here, every situation is different, so there's not a lot to say. However, the example in the notes it's a lot harder than usual. For intuition, it might help to look at the normalizing matrix \hat{R} in an intuitive way. If you write out each term, you will find that its i, j th element of \hat{R} is the sample correlation between two of the observed z vectors. Note also that D essentially plays the role of the \sqrt{n} in your more basic CLT. If the z vectors were not trending, then the d_j s would be on the order of \sqrt{n} .

2 Lent Term

2.1 2.1.1 Matrix Decomposition

I find the matrix decomposition non-intuitive. (Maybe there's something I don't know.) But an easy way to show the rank condition is by realizing that you don't change the rank of a matrix by multiplying it by a full rank matrix. This is similar to the trick above for deriving the inverse matrix formulas.

$$\begin{aligned}
 & \text{rank} \begin{pmatrix} \Pi' & I_K \\ W_{11} & W_{12} \end{pmatrix} \\
 = & \text{rank} \left[\begin{pmatrix} -C'B^{-1'} & I_K \\ W_{11} & W_{12} \end{pmatrix} \begin{pmatrix} I_G & 0_{G \times K} \\ C'B^{-1'} & I_K \end{pmatrix} \right] \\
 = & \text{rank} \left[\begin{pmatrix} 0 & I_K \\ W_{11} + W_{12}C'B^{-1'} & W_{12} \end{pmatrix} \begin{pmatrix} B' & 0 \\ 0 & I_K \end{pmatrix} \right] \\
 = & \text{rank} \begin{pmatrix} 0 & I_K \\ W_{11}B' + W_{12}C' & W_{12} \end{pmatrix} \\
 = & \text{rank} \begin{pmatrix} 0 & I_K \\ W_1A' & W_{12} \end{pmatrix}
 \end{aligned}$$

Note that, unlike the matrix decomposition, you can easily extend this trick to find necessary and sufficient conditions for estimation of the coefficients of several equations but not the whole system. A good exercise is to do it for the two-equation case.

2.2 Theorem 6

The end of this proof contains a fiddly error. From this argument, you can conclude that $\psi(\theta)$ is a constant function on θ on the parameterized path, not the entire neighborhood. This still means that θ is not L.I., of course.

2.3 Theorem 7

Again, this counterintuitive matrix decomposition is unnecessary if you use the same matrix tricks as above. Showing this is a good exercise. Note (and show) that

$$\Delta = I_G \otimes 2\Sigma B^{-1'}$$

Also, the (3,1) element of the second matrix in the decomposition should be $-\Delta$, not $-A$.

2.4 OLS and the Trace – section 3.4 and Theorem 25

It was not obvious to me, anyway, that $Q_2(\tilde{\Pi}, \Omega) \leq Q_2(\Pi, \Omega)$ if Π is unconstrained. Here is why:

Let Π be unconstrained, and let Ω be independent of θ . Then, substituting the reduced forms into the definition of the objective function and finding the first order conditions gives

$$\begin{aligned} \frac{\partial}{\partial \theta_i} n^{-1} \text{tr} \{V'V\Omega^{-1}\} &= \frac{\partial}{\partial \theta_i} n^{-1} \text{tr} \{(Y' - \Pi Z')(Y - Z\Pi')\Omega^{-1}\} \\ &= n^{-1} \text{tr} \{-2\Pi_i Z'(Y - Z\Pi')\Omega^{-1}\} = 0 \Rightarrow \\ \text{tr} \{\Pi_i Z'Y I_G \Omega^{-1}\} &= \text{tr} \{\Pi_i Z'Z\Pi'\Omega^{-1}\} \Rightarrow \\ \text{vec}'(\Pi_i) (Z'Y \otimes \Omega^{-1}) \text{vec}(I_G) &= \text{vec}'(\Pi_i) (Z'Z \otimes \Omega^{-1}) \text{vec}(\Pi') \Rightarrow \\ P' (Z'Y \otimes \Omega^{-1}) \text{vec}(I_G) &= P' (Z'Z \otimes \Omega^{-1}) \text{vec}(\Pi') \Rightarrow \\ \text{vec}(\Pi') &= [P' (Z'Z \otimes \Omega^{-1})]^{-1} P' (Z'Y \otimes \Omega^{-1}) \text{vec}(I_G) \\ &= ((Z'Z)^{-1} Z'Y \otimes I_G) \text{vec}(I_G) \\ &= \text{vec}(Y'Z(Z'Z)^{-1}) \\ &= \text{vec}(\tilde{\Pi}') \end{aligned}$$

This argument depends on the invertibility of P – that is, it depends on the fact that Π is unconstrained and identifiable. It also depends on the fact that θ

only parameterizes Π . But given those two things, the OLS estimator minimizes the trace of residuals no matter what the weighting matrix is.

This argument is also the key to understanding the "obvious" first step of Theorem 25.

2.5 Consistency of the 2SLS and 3SLS

I believe the logic is flawed in the assertion that

$$\|A - A_0\| \geq \epsilon \Rightarrow \|B\Pi_0 + C\|^2 \geq \delta_1$$

The book's reasoning would work if we knew that A were a continuous function of $B\Pi_0 + C$, but they argue instead that $B\Pi_0 + C$ is a continuous function of A , which gives the implication in the opposite direction that the proof needs. Essentially, one needs to ask: how do we know that, in our domain, $\|B\Pi_0 + C\|$ does not get arbitrarily close to zero away from the true value of θ without ever actually becoming 0? I would say something like this:

$A^* = \{\|A(\theta) - A(\theta_0)\|, \theta \in \Theta \setminus \{\theta : \|\theta - \theta_0\| \geq \epsilon\}\}$ is a compact set, since the domain Θ is assumed to be compact (and so Θ minus an open set is also compact), A is a continuous function on Θ (and so is the modulus of A minus a constant), and the domains of continuous functions defined on compact sets are compact. Since a compact set in \Re has a minimum, we can set $\delta_\epsilon = \min(A^*)$, and $\delta_\epsilon \neq 0$ since θ_0 is not in $\Theta \setminus \{\theta : \|\theta - \theta_0\| \geq \epsilon\}$. Thus $\inf_{\|\theta - \theta_0\| \geq \epsilon} (A(\theta) - A(\theta_0)) = \delta_\epsilon > 0$.

To be fair, I spent a long time talking with Prof. Hidalgo about this, and he insists that the book's argument is correct, but we were never able to convince each other, so I may just be missing something. And for what it's worth, I used this argument – not the book's – on the exam and they didn't fail me. (Who knows how many points they took off, though.) If you see something I don't, please email me.

2.6 Problem Set 4

Without giving anything away, let me just say that though this problem set is a lot of work the first time, it is not nearly as difficult and counterintuitive as our class teacher's notes made it seem. To navigate with ease, I recommend keeping these points in mind:

- Be careful with moving back and forth between the reduced form and structural form, and try to avoid doing it without a reason. (I think the book would be clearer if some of the derivations were re-written to follow

this principle – see my note about the derivative of $vec(\Pi)$ below.) Here, you are trying to get an answer in terms of the reduced form, but you have to think about the structural form because you are differentiating with respect to θ .

- Being fluent with the kronecker product and vec identities and derivatives of matrices and their inverses will be very important.
- When, in your derivation, you get to terms that are $O_p(0)$, get rid of them quickly and stop copying them separately every step of the way. You will save paper, time, and make fewer mistakes.
- Attack the problem one element of θ at a time.
- Keep track of where you're trying to get to, which is something that looks like linear and quadratic forms of P and W and don't worry about the slightly peculiar way the book writes the second derivative.
- In theorem 21, just before the conclusion, the book has two "non-conformal" matrices together in a partitioned matrix within a vec – that is, two matrices with different numbers of rows are concatenated horizontally. The formula is right if you just stack the columns of these two matrices proceeding left to right, but formally you might think it's confusing.
- Following these principles, the asymptotic distribution the 2SLS can be derived in a much simpler and more direct way than the solutions we were given in class. However, since that's a homework assignment, I'm not going to put it in here.

2.7 The derivative of $vec(\Pi)$ in section 5.0.1

I think the footnote here is a good example of unnecessary obfuscation. This derivative can be found very directly and intuitively:

$$\begin{aligned}
P_1 &= \frac{\partial}{\partial \theta_i} \text{vec}(\Pi') \\
&= \frac{\partial}{\partial \theta_i} \text{vec}(-C' B^{-1'}) \\
&= -\text{vec}(C'_i B^{-1'}) + \text{vec}(C' B^{-1} B'_i B^{-1'}) \\
&= -\text{vec}(C'_i B^{-1'} + \Pi' B'_i B^{-1'}) \\
&= -\text{vec}((C'_i + \Pi' B'_i) B^{-1'}) \\
&= -\text{vec}(\begin{pmatrix} \Pi' & I_k \end{pmatrix} A'_i B^{-1'}) \\
&= -\left(B^{-1} \otimes \begin{pmatrix} \Pi \\ I_k \end{pmatrix}' \right) \text{vec}(A'_i) \\
&= -\left(B^{-1} \otimes \begin{pmatrix} \Pi \\ I_k \end{pmatrix}' \right) \tilde{P}_i
\end{aligned}$$

2.8 Consistency of the Restricted LSE

Even if it seems a little trivial, I think an example might make this section clearer. Recall that, in general, A can be parameterized with fewer elements in θ than there are potential terms in the model. This is what constraints mean (see the A Few Notes On Different Estimators section for more about this). A very simple example of how this might be done with a linear model of the form given in 5.2.3 would be

$$\alpha_{11}y_1 + \alpha_{12}y_2 + \beta_{11}z_1 = \epsilon_1$$

$$\alpha_{22}y_2 + \beta_{22}z_2 = \epsilon_2$$

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \beta_{11} & \beta_{12} \\ \alpha_{21} & \alpha_{22} & \beta_{21} & \beta_{22} \end{pmatrix} \text{ (unconstrained)}$$

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \beta_{11} & 0 \\ 0 & \alpha_{22} & 0 & \beta_{22} \end{pmatrix} \text{ (constrained)}$$

$$\theta = \begin{pmatrix} a_{11} \\ a_{12} \\ b_{11} \\ a_{22} \\ b_{22} \end{pmatrix}$$

$$\text{vec}(A') = \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \beta_{11} \\ \beta_{12} \\ \alpha_{21} \\ \alpha_{22} \\ \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ b_{11} \\ a_{22} \\ b_{22} \end{pmatrix}$$

So we have

$$\theta = \begin{pmatrix} a_{11} \\ a_{12} \\ b_{11} \\ a_{22} \\ b_{22} \end{pmatrix}$$

$$F = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$f = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

which is a mapping from θ to A with the given form that imposes the necessary constraints. For the purpose of this section, they are restrict attention to cases where the constraints can be given a linear form.

Example 16, then, follows like so. Since in a multiple regression model we know that the coefficients of the y_i terms are not estimated, so the matrix F has G rows with all zeros, followed by K possible non-zero rows for each equation. And in the term $vec \begin{pmatrix} B_0^{-1} \Sigma_0 \\ 0 \end{pmatrix}$, there will be G rows of possibly non-zero values followed by K zeros, repeated as many times as you have equations (G times). The zeros in F' line up with the non-zero terms in $vec \begin{pmatrix} B_0^{-1} \Sigma_0 \\ 0 \end{pmatrix}$ and vice-versa, and the product is zero. (If this isn't clear, write it all out and you'll see what I mean.)

Example 17 is not quite as simple as this, but if you write it out you'll see that the argument is similar. Remember that a recursive system means that Σ_0 is diagonal!

2.9 Constraints and Parameterization

Here I'll talk a little about how I thought about the parameterization. Hopefully it will save you some of the time I spent staring at my shoes. Although it's not directly important by itself, and might seem a little fiddly, I found that thinking carefully about this stuff made a lot of otherwise mysterious proofs seem much clearer and more intuitive.

There is an important difference between the identification part (section 2) and the estimation part (section 3) in how they handle constraints which is explained in section 3.1. In section 2, the constraints are imposed explicitly with the matrix W . But in section 3, they are imposed implicitly by the mapping from θ to A and Σ and, hence, to Π and Ω , which are defined in terms of A and Σ . These constraints are imposed by minimizing the objective function with respect to θ only, which is lower-dimensional than A and Σ . The section I wrote above about Consistency of the Restricted LSE gives an example of how this might work in a simple case that could also be done in the style of section 2. Identification, in section 3, is proved more directly with more general assumptions.

In the framework of section 3, if something is "unconstrained", that just means all of its elements can be found as distinct elements of θ . For example, if Σ is unconstrained, then θ might contain first a parameterization of A and then $G \times G$ elements which are simply a list of the elements of Σ . (Of course, you could come up with a perverse way to do it some other way, but I think it's safe to just assume that you do it the simple way.) In this case, the derivative of Σ with respect to an element of θ that parameterized part of A would be zero, and the derivative with respect to the element of θ that parameterizes the $(i, j)^{th}$ element of Σ would be a matrix with all zeroes except for a one in the (i, j) place.

This can be confusing if Σ is unconstrained, and someone says that the derivative of Σ with respect to θ is zero. For example, in theorem 12, they say that " Σ is unconstrained and θ only parameterizes A ". What they really mean, in this case, is that we are really only interested in estimating the part of θ that parameterizes A , and that the derivative of Σ with respect to that part is zero.

Also, when you are taking derivatives of the reduced form terms Π and Ω with respect to θ , you should usually think of writing them first in terms of the structural forms, differentiating, and then taking the resulting expressions back into terms of the reduced form (if you can). This is because the relationship between the reduced form and θ is often implicitly defined via the structural form.

There are some caveats to this, though. Section 5.0.1 parameterizes Π and Ω directly from θ , for example. You could think of this as meaning that A and Σ are parameterized in such a way that Ω and Π end up parameterized sepa-

rately even though both depend on B (for example, B could be diagonal and Σ could be diagonal with elements proportional to the squares of the elements of B , and C could be parametrized so that the first row is proportional to the first element of B , and so on). Also, in the proof of the consistency of the estimators, the consistency of Π and Ω are shown, and the consistency of A and Σ follow directly from assumptions.

Despite this, I think it is less clear to think of θ as parameterizing Π and Ω directly and A and Σ being defined as functions of Π and Ω , because whatever those functions were would have to respect the definitions of the reduced form anyway.

Note also that it is also not necessarily true that Σ being unconstrained implies that Ω is unconstrained – consider, for example, if B was constrained to have only the $(1, 1)$ element be non-zero.

Don't forget that when you see α (which is usually taken to be a *vec* of the elements of A), it is not unconstrained. It, too, is a function of θ , just like A . (I see from some of my notes in the margin that I didn't understand that at first.)

Now, note that the PMLE estimator is set apart from the others in that it is the only one that explicitly parameterizes Σ . MLE, 2SLS, and 3SLS all provide estimates – even consistent estimates (under the necessary conditions, of course) – of Σ . That means that, as you get an infinite amount of data, the estimates will with probability approaching one become arbitrary close to respecting the constraints on Σ . However, this is not because the constraints are explicitly imposed, it is because the estimators are consistent and the true Σ obeys those constraints.

2.10 Miscellaneous Notes On the Estimators

As I write this now, some of this stuff looks kind of obvious, but I recall that it wasn't all obvious for me at the time I was first learning it, anyway, and seeing it all in once place might help you organize your thoughts more quickly than I did.

- The estimators all take the form of something that looks like a kind of least-squares (the trace of a term like $V'V$ is the sum of the squares of each equation's error term for each observation). (Of course, the PMLE also has a term that punishes a large variance, but it still also contains a sum of squared errors term.) The difference is that PMLE and the MDE look at the reduced form errors whereas the 2SLS and 3SLS look at the structural form errors (projected onto the space of the regressors to avoid endogeneity).

- The PMLE is the only one that parameterizes Σ directly, and so the only one that can accommodate explicit constraints on the covariance matrix.
- Although the PMLE and MDE are conceptually targeting the reduced form errors, they are often written in terms of the structural form variables because that make it easier to differentiate (see my section on parameterization above).
- The estimators differ in how they treat the covariance. The PMLE models it directly, the MDE uses OLS to get a consistent estimator (which is then fixed for the purposes of MDE estimation – that is, it does not depend on θ), the 2SLS does not use an estimate of the covariance at all, and the 3SLS uses an estimate implied from 2SLS (so that doesn't depend on θ , for the purposes of the 3SLS estimator, either).
- Only the 2SLS does not weight the sum of the squares of the errors by some kind of consistent estimate of the covariance matrix of those errors. For this reason, it alone has a different asymptotic distribution than the others even if Σ is unconstrained.