University of Illinois
Fall 2014

Department of Economics
Roger Koenker

Economics 508

Lecture 10

Introduction to Simultaneous Equation Econometric Models

1. Review of Linear "Seemingly Unrelated Regressions"

The simplest example of simultaneous equation models in econometrics is the model which Zellner labeled SUR and statisticians usually call just multivariate regression.

$$y_i = X_i \beta_i + u_i \qquad i = 1, \dots, m$$

where

$y_i \sim n$-vector of observed responses
$X_i \sim n \times p_i$ matrix of exogenous variables
$u_i \sim n$-vector of "errors"

A typical example would be a *system* of m demand equations in which X_i would be composed of prices and incomes and perhaps other commodity specific exogenous influences on demands. By exogenous in this preliminary setting we will simply mean that

$$EX_i^\top u_j = 0 \qquad i, j = 1, \dots, m.$$

which is the natural extension of the orthogonality condition underlying ordinary linear regression with a single response variable.

It is convenient to write the whole system of equations as

$$y = X\beta + u$$

which may be interpreted as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} X_1 & 0 & \cdots & \cdots & 0 \\ 0 & & & & \\ & & \ddots & & \\ 0 & & & & X_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$$

in which the equations have simply been stacked one on top of another. We will suppose that the full mn-vector, u, is normal with mean 0, and covariance matrix

$$Euu^\top = \Omega \otimes I_n = (\omega_{ij} \ \ I_n)$$

and we may then, immediately, write the optimal (unbiased) estimator of the parameter vector β as,

$$\hat{\beta} = (X^\top (\Omega \otimes I)^{-1} X)^{-1} X^\top (\Omega \otimes I)^{-1} y$$

where we note that $(\Omega \otimes I)^{-1} = \Omega^{-1} \otimes I$. Typically, Ω is unknown, but we may estimate it by $\Omega = (\hat{\omega}_{ij})$, with

$$\hat{\omega}_{ij} = \hat{u}_i^\top \hat{u}_j / n$$

where $\hat{u}_i, i = 1, \ldots, m$ are the n-vectors of residuals from any initial (consistent) estimate of the model, typically from an OLS fit to the individual equations.

An important observation is that there is no efficiency gain from the reweighting by $(\Omega \otimes I)^{-1}$ if $X = (I \otimes X_0)$. That is, if $X_i = X_0$ for all i as would be the case in some demand system contexts, we gain nothing from doing the system estimate over what is accomplished in the equation-by-equation OLS case. To see this write

$$(\Omega^{-1} \otimes I)(I \otimes X_0) = \Omega^{-1} \otimes X_0.$$

We are solving the equations in the weighted case

$$X^\top (\Omega \otimes I)^{-1} \hat{u} = 0$$

but if $X = (I \otimes X_0)$, this is equivalent to

$$(\Omega^{-1} \otimes X_0^\top)\hat{u}$$

but this is satisfied by assuring that

$$X_0^\top \hat{u}_i = 0 \qquad i = 1, \ldots, m$$

which are just the normal equations for the separate OLS regressions.

A useful introduction to maximum likelihood estimation of systems of equations may be provided by the SUR model. For this purpose it is convenient to stack the observations in "the opposite way" that is, to write

$$y_j = X_j \beta + u_j \qquad j = 1, \ldots, n$$

where

$$X_j = \begin{bmatrix} x_{j1} & 0 & \cdots & 0 \\ 0 & x_{j2} & \cdots & 0 \\ & & & \\ 0 & \cdots & & x_{jm} \end{bmatrix}$$

where x_{ji} is a p_i *row* vector. Now stacking the model we have,

$$y = X\beta + u$$

and now $u \sim \mathcal{N}(0, I \otimes \Omega)$. Note that, with this formulation

$$\begin{aligned} \hat{\beta} &= (X^\top (I \otimes \Omega^{-1})X)^{-1} X'(I \otimes \Omega^{-1})y \\ &= (\sum_{j=1}^{n} X_j^\top \Omega^{-1} X_j)^{-1} \sum_{j=1}^{n} X_j^\top \Omega^{-1} y_j \end{aligned}$$

The convenient aspect of this formulation is that we can view $u_j,\ j = 1, \ldots, n$ as independent realizations of an m-variate normal vector and thus the likelihood for the model may be written as,

$$\mathcal{L}(\beta, \Omega) = (2\pi)^{\frac{-mn}{2}} |\Omega|^{-\frac{n}{2}} \exp\{-\frac{1}{2} \sum_{j=1}^{n} u_j^\top \Omega^{-1} u_j\}$$

where implicitly we recognize that the u_j's are functions of the β vector. As usual it is more convenient to work with log likelihood,

$$\ell(\beta, \Omega) = K - \frac{n}{2} \log |\Omega| - \frac{1}{2} \sum u_j^\top \Omega^{-1} u_j$$

We have already seen how to estimate β in this model. We now consider two variants on estimation of Ω.

Case 1. Suppose that Ω is known up to a scalar, i.e., $\Omega = \omega\Omega_0$ with the matrix Ω_0 known.

Recall that $|\omega\Omega_0| = \omega^m|\Omega_0|$ so

$$\ell(\beta, \omega) = K - \frac{n}{2}(m \log \omega + \log |\Omega_0|) \quad - \frac{1}{2\omega}\sum u_j^\top \Omega_0^{-1} u_j$$

so

$$\frac{\partial l}{\partial \omega} = -\frac{nm}{2\omega} + \frac{1}{2\omega^2}\sum u_j^\top \Omega_0^{-1} u_j = 0$$

implies

$$\hat{\omega} = (mn)^{-1}\sum u_j^\top \Omega_0^{-1} u_j.$$

Case 2. If Ω is completely unknown, we simply differentiate with respect to Ω.

Now, from the Appendix,

$$\nabla_\Omega \ell = -\frac{n}{2}\Omega^{-1} + \frac{1}{2}\sum \Omega^{-1} u_j u_j^\top \Omega^{-1}$$

so

$$\hat{\Omega} = n^{-1}\sum u_j u_j^\top$$

which is the same formula suggested earlier in the lecture.

Now, concentrating the log likelihood as in the single equation case we may simplify the last term,

$$\sum \text{ tr } (u_j^\top (\sum u_j u_j^\top)^{-1} u_j) = \sum \text{ tr } u_j u_j^\top (\sum u_j u_j^\top)^{-1} = mn$$

so for purposes of computing likelihood ratios or SIC numbers we have

$$\ell(\hat{\beta}, \hat{\Omega}) = K^* - \frac{n}{2}\log |\hat{\Omega}|$$

where K^* is a constant independent of the data. Thus, maximizing the likelihood, or log-likelihood is the same as minimizing the determinant of $\hat{\Omega}$, which is sometimes called the generalized variance.

2. Introduction to Vector Autoregressive Models

An important class of models in time-series which draw upon the ideas of SUR models are the so called VAR models. Consider an m-vector y_t observed at time t and a model

$$y_t = \mu + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + u_t$$

Again exploiting the lag operator notation we may write this as

$$A(L) y_t = \mu + u_t$$

where

$$A(L) = I - A_1 L - A_2 L^2 - \ldots - A_p L^p.$$

Again, stability is crucial determined by the characteristic equation*,

$$|A(z)| = 0.$$

If the roots of this equation lie outside the unit circle, then all is well, if some roots lie on the unit circle, then it is useful to reformulate the model in the error correction form

$$\Delta y_t = \mu + B_1 \Delta y_{t-1} + B_2 \Delta y_{t-2} + B_{p-1} \Delta y_{t-p+1} - \Pi y_{t-1} + u_t \tag{0}$$

where we have

$$\Pi = A(1) = I - A_1 - \ldots - A_p$$

and has rank less than m. Note that this is analogous to the reformulation of the univariate model leading to ADF test. We then factor $\Pi = AB$ which is singular into pieces, A and B that have rank r, the nonsingular part, $m - r$, the singular part, respectively, and this leads to the theory of cointegrated time series, a topic which is dealt with in some depth in our graduate time series course. I'll provide a somewhat sketchy introduction at the end of this lecture. The integer r is called the rank of the cointegrating relationship and denotes the number of linear combinations of the original m variables that are stationary.

3. Impulse Response Functions, Again

Since we have a somewhat different setting than our single equation demand model, it is worth revisiting the question "what is an IRF for a VAR?" In the VAR context we have no exogenous variables which might be regarded as candidates for a permanent policy shock of the type we have already discussed.

However, we can still ask what would be the path of the system if it were in equilibrium and was then "shocked" by a permanent increase in one of the error realizations. So we are really asking what happens to the whole system of equations, how does it evolve after encountering a once and for all increase in one element of the error vector u_t. Formally, we have the same problem except that now we have matrices everywhere we used to have scalars.

If the model is stable in the sense we have already described, we can "invert" the VAR representation and put the model in the MA form,

$$y_t = m + A(L)^{-1} u_t$$

where $A(L)^{-1} u_t$ is interpretable in much the same way that we interpreted

$$D(L) x_t = A(L)^{-1} B(L) x_t$$

in the earlier, simpler, models. To illustrate, it may be helpful to consider the example,

$$A(L) = I - AL$$

In this case the invertible MA representation would have

$$(I - AL)^{-1} = I + AL + A^2 L^2 + \ldots$$

Note that as in the simple case we can verify this directly. Obviously we require that the right hand side converge, for this to make any sense. The MA or impulse response formulation of the model has some inherent ambiguity in the typical case of correlated errors. The underlying thought experiment is rather implausible in this case and there

[*] Note that this characteristic equation now involves the determinant of a matrix, not simply an ordinary polynomial, but the principle is the same as before.

has been considerable discussion about various schemes to orthogonalize the errors, but these “solutions” introduce new problems having to do with the nonuniqueness of the orthogonalization.

There are two common procedures for testing for cointegration, one introduced by Engle and Granger (1987), the other by Johansen (1988). We will describe both very succinctly. Consider the problem posed in PS3 of testing for the cointegration of two series $\{x_t, y_t\}$. If we knew the coefficients of the cointegrating relationship, i.e., if we hypothesized, for example,

$$z_t = y_t - \alpha - \beta x_t$$

was stationary then the situation would be relatively simple we would simply apply the Dickey-Fuller as the augmented Dickey-Fuller test to the new variable z_t. If we *reject* the null hypothesis that the series z_t *has* a unit root, having already concluded that the hypotheses that x_t and y_t themselves *cannot* be rejected, then we may conclude that there is evidence for the cointegration of $\{x_t, y_t\}$

It may seem very implausible that we might “know” α, β, however in some examples this isn’t so strange. Often theory might suggest that $(\alpha, \beta) = (0, 1)$ is reasonable. But what should we do if we don’t have any *a priori* notion about (α, β)?

Fortunately, the answer to this question, at least from a pragmatic point of view is quite simple. As a first step we estimate the parameters using the usual least squares procedure, and then proceed as before using $\hat{z}_t = y_t - \hat{\alpha} - \hat{\beta} x_t$. The only difference is that some adjustment in the original DF critical values is necessary. For the problem set, these new critical values are provided in Table B.9 from Hamilton (1994). There are additional complications due to trends, but I will defer these to the complete treatment of these matters in 574, our time-series course.

In the case that we have more than two series the situation is a bit more complicated. An elegant general approach is provided by the canonical correlation methods of Johansen (1991). Johansen’s approach employs two sets of auxiliary regressions. Returning to the prior matrix notation, write

$$\Delta y_t = \hat{\delta}_0 + \hat{\delta}_1 \Delta y_{t-1} + \ldots + \hat{\delta}_{p-1} \Delta y_{t-p+1} + \hat{u}_t$$

and

$$y_{t-1} = \hat{\theta}_0 + \hat{\theta}_1 \Delta y_{t-1} + \ldots + \hat{\theta}_{p-1} \Delta y_{t-p+1} + \hat{v}_t$$

These two equations can be viewed as providing the first stage of the partial residual plot analysis of the error correction model introduced earlier, having estimated them, we can explore the rank of the Π matrix in that equation. It turns out that the likelihood for the original VAR with $Eu_t u_s' = \Omega$ for $t = s$ and $= 0$ otherwise, is

$$\ell = -\frac{nm}{2} \log(2\pi) - \frac{nm}{2} - \frac{n}{2} \log|\hat{\Sigma}_{uu}| - \frac{n}{2} \sum_{i=1}^{r} \log(1 - \hat{\lambda}_i)$$

where $\hat{\lambda}_i$ denote ordered eigenvalues $\hat{\lambda}_1 > \lambda_2 > \ldots > \hat{\lambda}_m$ of the matrix $\hat{\Sigma}_{vv}^{-1} \hat{\Sigma}_{vu} \hat{\Sigma}_{uu}^{-1} \hat{\Sigma}_{uv}$, and r, as above, denotes the rank of the cointegrating relationship, that is the rank of the matrix Π. An LR test based on this approach and intended to test rank r against rank

m employs the test statistic,

$$2(\ell_1 - \ell_0) = -n \sum_{i=r+1}^{m} \log(1 - \hat{\lambda}_i)$$

For another example, in the simple bivariate example of PS3, our test would correspond to testing the null hypothesis of $r = 0$ cointegrating vectors against the alternative of $r = 1$ cointegrating vectors. Here $m = 2$ so we have the test statistic

$$2(\ell_1 - \ell_0) = -n \log(1 - \lambda_1)$$

From Table B.10 of Hamilton (1994) we find that the critical value of this test is 3.84 at 5%. A reasonably complete derivation of these expressions is provided by the semi-historical discussion in the concluding sections.

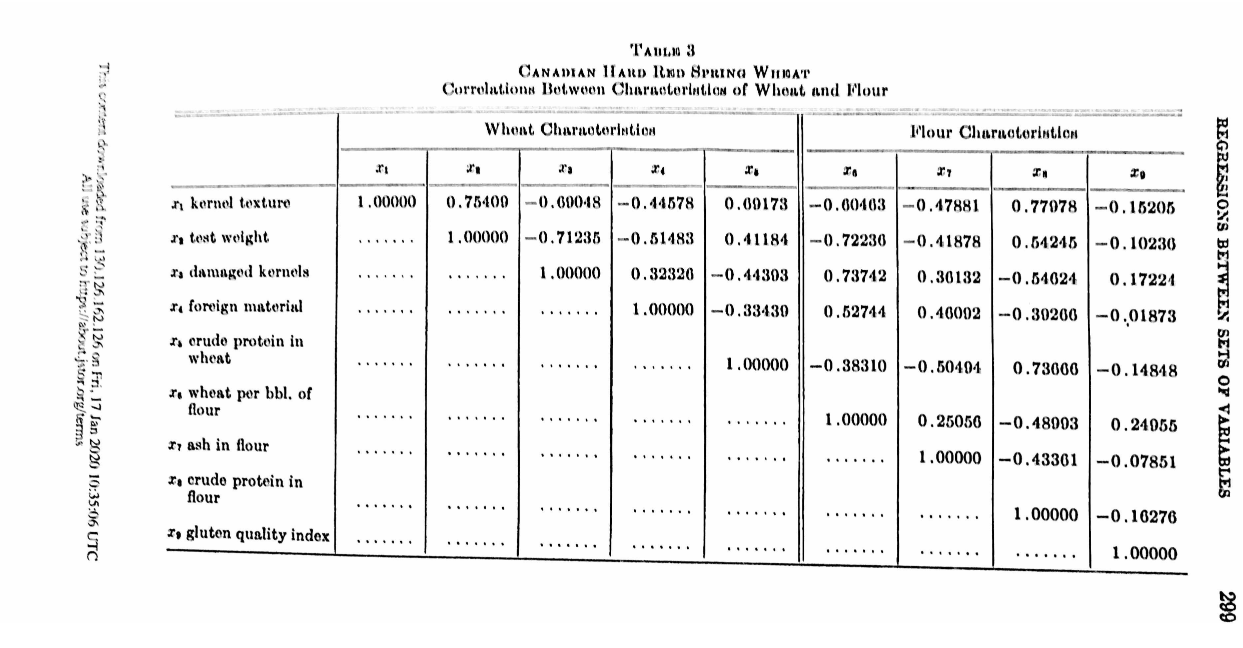

Table 3
Canadian Hard Red Spring Wheat
Correlations Between Characteristics of Wheat and Flour

	Wheat Characteristics					Flour Characteristics			
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1 kernel texture	1.00000	0.75409	−0.69048	−0.44578	0.69173	−0.60463	−0.47881	0.77078	−0.15205
x_2 test weight	....	1.00000	−0.71235	−0.51483	0.41184	−0.72236	−0.41878	0.54245	−0.10236
x_3 damaged kernels	....	....	1.00000	0.32326	−0.44393	0.73742	0.36132	−0.54624	0.17224
x_4 foreign material	....	....	....	1.00000	−0.33439	0.52744	0.46092	−0.39266	−0.01873
x_5 crude protein in wheat	....	....	....	....	1.00000	−0.38310	−0.50494	0.73666	−0.14848
x_6 wheat per bbl. of flour	....	....	....	....	....	1.00000	0.25056	−0.48993	0.24955
x_7 ash in flour	....	....	....	....	....	....	1.00000	−0.43361	−0.07851
x_8 crude protein in flour	....	....	....	....	....	....	....	1.00000	−0.16276
x_9 gluten quality index	....	....	....	....	....	....	....	....	1.00000

REGRESSIONS BETWEEN SETS OF VARIABLES 299

Figure 1. Source: Waugh (1942)

4. Canonical Correlation

Regression generalizes the notion of simple bivariate correlation by finding the linear combination of a vector of covariates that is most highly correlated with the response variable, that is it finds the linear combination of x's that maximizes the correlation of y and $\hat{y}$.

$$R^2 = 1 - \frac{\hat{\sigma}^2_{y|x}}{\hat{\sigma}^2_y} = \rho^2(y, \hat{y}).$$

This is just the squared cosine of the angle between y and $\hat{y}$.

Hotelling (1935) generalized this notion further to handle a vector of response variables. Suppose we have, m y's and p x's? Consider two arbitrary linear combinations $\alpha^\top y$ and $\beta^\top x$ and suppose for convenience α and β could be chosen so that they yield unit variance. We would like to choose α and β so that they maximized the correlation between $\alpha^\top y$ and $\beta^\top x$. How? Let's write: $\text{Cov}(\alpha^\top y, \beta^\top x) = \alpha^\top \Sigma_{xy}\beta$, $V(\alpha^\top y) = \alpha^\top \Sigma_{yy}\alpha$ and $V(\beta^\top x) = \beta^\top \Sigma_{xx}\beta$. We want to maximize the Lagrangean expression

$$\alpha^\top \Sigma_{xy}\beta - \frac{\lambda_1}{2}\alpha^\top \Sigma_{yy}\alpha - \frac{\lambda_2}{2}\beta^\top \Sigma_{xx}\beta$$

so we have the first order conditions

$$\begin{aligned} \Sigma_{xy}\beta &= \lambda_1 \Sigma_{yy}\alpha \\ \Sigma_{xy}\alpha &= \lambda_2 \Sigma_{xx}\beta \end{aligned}$$

multiplying through by α and β respectively gives $\lambda_1 = \alpha^\top \Sigma_{xy}\beta$ and $\lambda_2 = \alpha^\top \Sigma_{xy}\beta$, so $\lambda_1 = \lambda_2 = \rho$. Now multiply the second equation by $\Sigma_{xy}\Sigma_{xx}^{-1}$

$$\Sigma_{xy}\Sigma_{xx}^{-1}\Sigma_{xy}\alpha = \rho\Sigma_{xy}\beta$$

and multiply the first by ρ

$$\rho^2\Sigma_{yy}\alpha = \rho\Sigma_{xy}\beta$$

and subtract to obtain the eigenvalue problem,

$$\Sigma_{xy}\Sigma_{xx}^{-1}\Sigma_{xy}\alpha - \rho^2\Sigma_{yy}\alpha = 0.$$

The usual eigenvalue problem $(A - \lambda I)x = 0$ implies $Ax = \lambda x$ so the vector x is unaltered in direction by the transformation A except in length. Our problem can be reformulated in this way by writing

$$(\Sigma_{yy}^{-1}\Sigma_{xy}\Sigma_{xx}^{-1}\Sigma_{xy} - \rho^2 I)\alpha = 0.$$

Now of course there is more than one eigenvalue – there are m of them, where m is the dimension of the response y. They can be ordered and for this problem they are called the canonical correlations and the corresponding eigenvectors are called the canonical variables. The latter are constructed in the following way: Given a pair (ρ^2, α) satisfying (3) we can find an associated β by simply regressing $\alpha^\top y$ on x, thereby constructing all of the triples, (ρ^2, α, β).

To illustrate this we can revisit an example introduced by Waugh (1942). Yes, that Waugh. He writes: "Professor Hotelling's paper, should be widely known and his method used by practical statisticians. Yet, few practical statisticians seem to know of the paper, and perhaps those few are inclined to regard it as a mathematical curiosity rather than an important and useful method of analyzing concrete problems." The second of Waugh's examples concerns the quality of wheat and how it influences the quality of flour it produces. He uses data from 136 export shipments of hard spring wheat. The correlation matrix of the data is shown in the following table. There are five variables for wheat quality, and four for quality of the flour. In 1942 it involved a non-trivial amount of effort to compute the canonical correlations and Waugh adopts some clever iterative tricks to get an approximation, but it is trivial to do so now in R. Applying the machinery introduced above, we get: $\rho = c(0.910, 0.650, 0.269, 0.169)$.

5. Reduced Rank Regression

Canonical correlation is intimately connected to several important maximum likelihood problems in classical econometrics. The most basic version of these problems is represented by the reduced rank regression procedure of Anderson (1951). Consider the model,

$$Y_t = AB^\top X_t + DZ_t + u_t$$

where $u \sim \mathcal{N}(0, \Omega)$. The log likelihood is given by

$$\ell(A, B, D, \Omega) = K - \frac{n}{2} \log |\Omega| - \frac{1}{2} \sum u_t^\top \Omega^{-1} u_t.$$

As usual we can concentrate the likelihood,

$$\ell(A, B, \Omega) = -\frac{n}{2} \log |\Omega| - \frac{1}{2} \Sigma \tilde{u}_t^\top \Omega^{-1} \tilde{u}_t$$

where $\tilde{u}_t = \tilde{Y}_t - AB^\top \tilde{x}_t, \quad \tilde{Y} = (I - P_Z)Y, \quad \tilde{X}_t = (I - P_Z)X$. For fixed B we can easily estimate A in this reduced model

$$\tilde{Y}_t = AB^\top \tilde{X}_t + \tilde{u}_t$$

and would obtain,

$$\hat{A}(B) = (B^\top \tilde{X}^\top \tilde{X} B)^{-1} B^\top \tilde{X}^\top \tilde{Y}$$

and

$$\hat{\Omega}(B) = \Sigma_{\tilde{Y}\tilde{Y}} - \Sigma_{\tilde{X}\tilde{Y}} B (B^\top \Sigma_{\tilde{X}\tilde{X}}^{-1} B)^{-1} B^\top \Sigma_{\tilde{X}\tilde{Y}} = \Sigma_{\tilde{Y}\tilde{Y}} - \hat{A}(B)(B^\top \Sigma_{\tilde{X}\tilde{X}}^{-1} B)^{-1} \hat{A}(B)$$

where $\Sigma_{\tilde{Y}\tilde{Y}} = \tilde{Y}^\top \tilde{Y}/n$, etc. As usual the likelihood after substitution looks like this:

$$\ell(\hat{A}, \hat{B}, \hat{\Omega}) = K - \frac{n}{2} \log |\hat{\Omega}|$$

where K is independent of the data. So we are simply trying to minimize the generalized variance represented by the determinant of Ω.

Now we need some trickery involving determinants of partitioned matrices. We start, following Johansen, with the identity

$$\begin{vmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{vmatrix} = |\Sigma_{00}| \; |\Sigma_{11} - \Sigma_{10} \Sigma_{00}^{-1} \Sigma_{01}| = |\Sigma_{11}| \; |\Sigma_{00} - \Sigma_{01} \Sigma_{11}^{-1} \Sigma_{10}|.$$

Thus,

$$\begin{aligned} \begin{vmatrix} \Sigma_{00} & \Sigma_{01} B \\ B^\top \Sigma_{10} & B^\top \Sigma_{11} B \end{vmatrix} &= |\Sigma_{00}| \; |B^\top (\Sigma_{11} - \Sigma_{10} \Sigma_{00}^{-1} \Sigma_{01}) B| \\ &= |B^\top \Sigma_{11} B| \; |\Sigma_{00} - \Sigma_{01} B (B^\top \Sigma_{11}^{-1} B) B^\top \Sigma_{01}|, \end{aligned}$$

and therefore,

$$|\Sigma_{00} - \Sigma_{01} B(B^\top(\Sigma_{11}^{-1} B)^{-1} B^\top \Sigma_{01}| = |\Sigma_{00}| \ |B^\top(\Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01})B| \ / \ |B^\top \Sigma_{11} B|.$$

Translating this expression back into our likelihood notation we have

$$(*) \qquad |\hat{\Omega}| = |\Sigma_{\tilde{Y}\tilde{Y}}| \ |B^\top(\Sigma_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}\tilde{Y}}\Sigma_{\tilde{Y}\tilde{Y}}^{-1}\Sigma_{\tilde{X}\tilde{Y}})B| \ / \ |B^\top \Sigma_{\tilde{X}\tilde{X}} B|.$$

Recall that we are still trying to maximize $-\log|\hat{\Omega}|$ with respect to B. How do this?

Consider a simpler version of a similar problem. Suppose A is a symmetric positive semi-definite matrix and we want to solve:

$$\max_x \ \frac{x^\top A x}{x^\top x}$$

We can write $A = P^\top \Lambda P = \Sigma \lambda_i p_i p_i^\top$ where $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ are the eigenvalues of A and P is the matrix of corresponding eigenvectors. P constitutes a basis for the space $\mathcal{R}^n$ so any x can be written as

$$x = P\alpha$$

so our problem becomes,

$$\max_\alpha \ \sum \lambda_i \alpha_i^2 / \sum \alpha_i^2$$

which is accomplished by letting $\alpha = (1, 0, \cdots, 0)$ giving λ_1 as the maximum.

Generalizing slightly, consider for A symmetric positive semi-definite and B symmetric positive definite

$$\max_x \ x^\top A x / x^\top B x.$$

Write $B = C^\top C$ as the Cholesky decomposition (matrix square root) of B, set $y = Cx$ to get

$$\max_y \ y^\top C^{-1} A C^{-1} y / y^\top y$$

so the solution is found by choosing the largest eigenvalue of the matrix $C^{-1}AC^{-1}$. This generalized eigenvalue problem can be posed as finding roots of

$$|\lambda A - B| = 0.$$

This all generalizes to finding multiple roots and multiple eigenvectors. Let X be an $n \times p$ matrix and consider

$$\max_X |X^\top A X| / |X^\top B X|.$$

Similar arguments lead to choosing X to represent to the matrix of eigenvectors corresponding to the p largest eigenvalues of $C^{-1}AC^{-1}$ where again $B = C^\top C$, by solving

$$|\lambda A - B|.$$

Now, finally, we are ready to get back to the problem of maximizing $-|\hat{\Omega}|$. This requires solving the eigenvalue problem

$$|\rho \Sigma_{\tilde{X}\tilde{X}} - (\Sigma_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{Y}\tilde{X}}\Sigma_{\tilde{Y}\tilde{Y}}^{-1}\Sigma_{\tilde{Y}\tilde{X}}| = 0.$$

or setting $\rho = (1 - \lambda)$

$$|\lambda \Sigma_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{Y}\tilde{X}}\Sigma_{\tilde{Y}\tilde{Y}}^{-1}\Sigma_{\tilde{Y}\tilde{X}}| = 0.$$

And thus the likelihood can be expressed as

$$\ell(A,B,\Omega) = K - \frac{n}{2}\log|\Sigma_{\tilde{X}\tilde{X}}| - \frac{n}{2}\log\Sigma_{i=1}^{r}\log(1-\hat{\lambda}_i).$$

Appendix On Matrix Differentiation

Three basic results on matrix differentiation may be recalled here.

$$(i) \qquad \frac{\partial \log|A|}{\partial A} = (A')^{-1} \qquad \text{if } |A|>0$$

$$(ii) \qquad \frac{\partial x'Ax}{\partial A} = xx'$$

$$(iii) \qquad \frac{\partial y'A^{-1}y}{\partial A} = -A^{-1}yy'A^{-1}$$

It is perhaps useful first to verify that these formulae work in the "A scalar" case. We provide sketchy arguments for each of the general results below.

(i): This follows from the fact that the determinant can be expanded as

$$|A| = \sum_{i=1}^{n} a_{ij}A_{ij}$$

where a_{ij} is the ij^{th} element of A and $(-1)^{i+j}A_{ij}$ is the ij^{th} *cofactor* of A, that is the determinant of the matrix A with the i^{th} row and j^{th} column deleted. Thus, the derivative of $|A|$ with respect to a_{ij} is just A_{ij} which is $|A|$ times the ji^{th} element of A^{-1}. Thus $\partial|A|/\partial A = |A|(A^{-1})'$, and thus by the chain rule, we have (i). Note $(A')^{-1} = (A^{-1})'$, and the transpose is usually irrelevant since A is symmetric in (most) applications.

(ii): Write $x'Ax = \sum a_{ij}x_ix_j$, so

$$\frac{\partial x'Ax}{\partial A} = \left(\frac{\partial x'Ax}{\partial a_{ij}}\right) = (x_ix_j)$$

(iii): To see this write

$$\frac{\partial x'A^{-1}x}{\partial A} = x'\frac{\partial A^{-1}}{\partial A}x$$

and differentiate the identity $AA^{-1} = I$ to obtain,

$$0 = \frac{\partial A}{\partial a_{ij}}A^{-1} + A\frac{\partial A^{-1}}{\partial a_{ij}}$$

so

$$\frac{\partial A^{-1}}{\partial a_{ij}} = -A^{-1}\frac{\partial A}{\partial a_{ij}}A^{-1}$$

where $\partial A/\partial a_{ij} = e_ie_j'$ is a matrix with ij^{th} element 1 and the rest zeros. Thus

$$\frac{\partial x'A^{-1}x}{\partial a_{ij}} = -x'A^{-1}e_ie_j'A^{-1}x = -e_i'A^{-1}xx'A^{-1}e_j$$

and (iii) follows by arranging the elements in matrix form.