

# SIMULTANEOUS EQUATION SYSTEMS

Myung Seo<sup>1</sup>

March 15, 2012

<sup>1</sup>Office Hour: 1:30PM - 2:30PM on Wed at S580

# Contents

<b>1</b>	<b>PRELIMINARIES.</b>	<b>5</b>
1.1	LINEAR (IN VARIABLES) MODELS . . . . .	7
1.2	NONLINEAR SEM . . . . .	16
<b>2</b>	<b>IDENTIFICATION</b>	<b>21</b>
2.1	BASIC DEFINITIONS AND EXAMPLES . . . . .	21
2.2	IDENTIFICATION OF LINEAR SEM . . . . .	35
2.2.1	IDENTIFICATION OF A SINGLE EQUATION . . . . .	40
2.3	NONLINEARITY AND LOCAL IDENTIFICATION . . . . .	45
2.4	IDENTIFIABILITY AND ASYMPTOTIC THEORY . . . . .	64
<b>3</b>	<b>ESTIMATION OF LINEAR SEM</b>	<b>69</b>
3.1	PARAMETERIZATIONS . . . . .	69
3.2	GAUSSIAN PSEUDO-MAXIMUM LIKELIHOOD ESTIMATOR (PMLE) . . . . .	73

3.3	MINIMUM DISTANCE ESTIMATOR (MDE) (NLLS)	80
3.4	TWO STAGE AND THREE STAGE LEAST SQUARES	82
3.5	INDIRECT LEAST SQUARES (ILSE). RELATIONSHIP BETWEEN DIFFERENT ESTIMATORS	89
3.6	SINGLE EQUATION AND SUBSYSTEM ESTIMATION	95
<b>4</b>	<b>CONSISTENCY OF THE ESTIMATORS</b>	<b>102</b>
4.1	CONSISTENCY OF THE PMLE	103
4.2	CONSISTENCY OF THE MDE	112
4.3	CONSISTENCY OF THE 2SLSE AND 3SLSE	114
<b>5</b>	<b>ASYMPTOTIC NORMALITY OF THE ESTIMATORS</b>	<b>123</b>
5.1	PMLE	125
5.1.1	WHEN $\Pi$ AND $\Omega$ ARE FUNCTIONALLY UNRE- LATED	131
5.2	ASYMPTOTIC EQUIVALENCE OF ESTIMATES	136
5.3	DELTA METHOD	150

5.4	<b>LAGRANGE MULTIPLIER METHODS . . . . .</b>	<b>153</b>
5.4.1	<b>CONSISTENCY OF RESTRICTED LSE OF <math>A_0</math> . . . . .</b>	<b>158</b>
<b>6</b>	<b>ESTIMATION OF NONLINEAR SEM AND TRANSFORMA- TION MODELS . . . . .</b>	<b>164</b>
6.1	<b>MAXIMUM LIKELIHOOD ESTIMATOR (GAUSSIAN) . . . . .</b>	<b>165</b>
6.2	<b>INSTRUMENTAL VARIABLES. GMM ESTIMATES . . . . .</b>	<b>169</b>
6.2.1	<b>FEASIBLE OPTIMAL IV ESTIMATOR . . . . .</b>	<b>176</b>
6.2.2	<b>Linear SEM . . . . .</b>	<b>184</b>
<b>7</b>	<b>HYPOTHESIS TESTING . . . . .</b>	<b>188</b>
7.1	<b>Review of Classical Hypothesis Testing . . . . .</b>	<b>190</b>
7.2	<b>WALD TEST (Generalized) . . . . .</b>	<b>199</b>
7.2.1	<b>CONSIDER EXTREMUM ESTIMATES . . . . .</b>	<b>219</b>
7.3	<b>THE LANGRANGE MULTIPLIER TEST (SCORE TEST) . . . . .</b>	<b>226</b>
7.4	<b>THE PSEUDO LIKELIHOOD RATIO TEST . . . . .</b>	<b>238</b>

<b>8</b>	<b>Inference under More General Conditions</b>	<b>257</b>
<b>9</b>	<b>ASYMPTOTICS WHEN THE TRUE VALUE OF THE PARAMETER IS AT THE BOUNDARY</b>	<b>260</b>
9.1	NONLINEAR MODELS . . . . .	270

# 1 PRELIMINARIES.

- A simultaneous equations model (SEM) concerns a  $(G + K) \times 1$  vector  $x_i$  of economic variables,  $i = 1, 2, \dots$ , where

$$x_i = \begin{pmatrix} y_i \\ z_i \end{pmatrix} \begin{array}{l} \rightarrow G \times 1 \\ \rightarrow K \times 1, \end{array}$$

- $y_i$  is a set of *endogenous* variables (dependent variable) and
- $z_i$  is a collection of *predetermined* variables (regressors)
- The predetermined variables consist of *exogenous* variables and lagged endogenous variables.
- Typically,  $G > 1$  and the endogenous variables in  $y_i$  are simultaneously determined given  $z_i$  and other unobservable variables  $u_i$ , as predicted by an economic theory, that is,

$$u(x_i, \theta_0) = u_i \rightarrow M \times 1,$$

where the functional form  $u(\cdot, \cdot)$  is known or given by an economic theory and the true parameter value  $\theta_0$  is unknown.

- Leading economic examples are the demand-supply system and structural vector autoregression for macro economic models.<sup>1</sup>

---

<sup>1</sup>Read discussion articles in *Journal of Economic Perspectives* (2010, vol 24, No 2) for different views on econometric models, in particular, the one by Sims on SEMs.

## 1.1 LINEAR (IN VARIABLES) MODELS

The majority of SEMs take the linear form,

$$Ax_i = u_i, \quad (1.1)$$

In (1.1),  $A$  is a  $M \times (G + K)$  matrix of parameters, so that  $u_i$  is  $(M \times 1)$ -vector of error terms.

depending on the model, we shall require

$$\text{Cov}(z_i; u_i) = 0 \quad (1.2)$$

$$E[u_i | z_i] = 0 \quad (1.3)$$

$$(u_i \text{ and } z_i) \text{ are independent.} \quad (1.4)$$

We have that

$$(1.4) \Rightarrow (1.3) \Rightarrow (1.2).$$

**Definition 1** Equation (1.1) is called *Structural Form*, whereas the matrix of parameters “ $A$ ” is called *Structural Form parameters*. Moreover,  $u_i$  is the *Structural Form disturbance vector*.

We will always assume that

$$E(u_i u_i') = \Sigma$$

is a semi positive definite finite matrix. That is, the error term  $u_i$  has finite second moments. Let

$$A = \begin{pmatrix} B & C \\ (M \times G) & (M \times K) \end{pmatrix}.$$

Then (1.1) can be written as

$$\begin{aligned} Ax_i &= (B; C) \begin{pmatrix} y_i \\ z_i \end{pmatrix} \\ &= By_i + Cz_i \\ &= u_i. \end{aligned} \tag{1.5}$$

From (1.5) we observe that we have a system with  $M$  equations.

**Definition 2** *If  $M < G$  we say that the system of equations is incomplete, whereas if  $M = G$ , then we say that the system of equations is complete.*

Herewith, we shall assume that  $|B| \neq 0$ , so that  $B^{-1}$  exists, and thus (1.5) can be written as

$$\begin{aligned} B^{-1} [By_i + Cz_i] &= B^{-1}u_i \\ y_i - \Pi z_i &= v_i \end{aligned}$$

or

$$y_i = \Pi z_i + v_i. \tag{1.6}$$

**Definition 3** *The matrix  $\Pi = -B^{-1}C$  is called the reduced form parameters, and (1.6) is known as the reduced form. Moreover,  $v_i = B^{-1}u_i$  is the reduced form disturbance vector.*

Because  $E(u_i u_i') = \Sigma$ , we have that

$$E(v_i v_i') = B^{-1} \Sigma (B')^{-1} = \Omega.$$

Our questions are:

- (i) Can we estimate  $A$ ?
- (ii) What are the statistical properties of such an estimate of  $A$ , say  $\hat{A}$ ?
- (iii) How can we easily obtain  $\hat{A}$ ?

- Sims (1980)'s seminal work: Structural VAR model

$$\begin{array}{llll}
 m & \text{money} & m & = \varepsilon_m \\
 y/p & \text{real GNP} & y/p & = \beta_{21}m + \varepsilon_{y/p} \\
 u & \text{unemployment} & u & = \beta_{31}m + \beta_{32}y/p + \varepsilon_u \\
 w & \text{wage level} & w & = \beta_{41}m + \beta_{42}y/p + \beta_{43}u + \varepsilon_w \\
 p & \text{price level} & p & = \beta_{51}m + \beta_{52}y/p + \beta_{53}u + \beta_{54}w + \varepsilon_p \\
 pm & \text{import price} & pm & = \beta_{61}m + \beta_{62}y/p + \beta_{63}u + \beta_{64}w + \beta_{65}p + \varepsilon_{pm}
 \end{array}$$

- small number of equations
- no exogenous variables

- Blanchard (1989)

variables		innovations	
$y$	output	$\varepsilon_d$	aggregate demand
$u$	unemployment rate	$\varepsilon_\theta$	labor supply and technology
$p$	price level	$\varepsilon_p$	price
$w$	wage level	$\varepsilon_w$	wage level
$m$	nominal money	$\varepsilon_m$	nominal money

---

equations

---

aggregate demand	$y = \delta_{12}\varepsilon_\theta + \varepsilon_d$
Okun's law	$u = \gamma_{21}y + \varepsilon_\theta$
price setting	$p = \gamma_{34}w + \gamma_{31}y + \delta_{32}\varepsilon_\theta + \varepsilon_p$
wage setting	$w = \gamma_{43}p + \gamma_{42}u + \delta_{42}\varepsilon_\theta + \varepsilon_w$
money rule	$m = \gamma_{51}y + \gamma_{52}u + \gamma_{53}p + \gamma_{54}w + \varepsilon_m$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -\gamma_{21} & 1 & 0 & 1 & 0 \\ -\gamma_{31} & 0 & 1 & -\gamma_{34} & 0 \\ 0 & -\gamma_{42} & -\gamma_{43} & 1 & 0 \\ -\gamma_{51} & -\gamma_{52} & -\gamma_{53} & -\gamma_{54} & 1 \end{pmatrix} \begin{pmatrix} y \\ u \\ p \\ w \\ m \end{pmatrix} = \begin{pmatrix} 1 & \delta_{12} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & \delta_{32} & 1 & 0 & 0 \\ 0 & \delta_{42} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_d \\ \varepsilon_\theta \\ \varepsilon_p \\ \varepsilon_w \\ \varepsilon_m \end{pmatrix}$$

### **Identities.**

That is,

$$Ax_i = u_i,$$

where

$$\begin{aligned} u_i &= \begin{pmatrix} u_{1i} \\ 0 \end{pmatrix} \rightarrow (G_1 \times 1) \\ &\quad \rightarrow ((G - G_1) \times 1) \\ A &= \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \rightarrow (G_1 \times (K + G)) \\ &\quad \rightarrow (G - G_1) \times (K + G) \\ A &= \begin{pmatrix} A_1 = (B_{11}; B_{12}; C_1) \\ A_2 = (B_{21}; B_{22}; C_2) \end{pmatrix}. \end{aligned}$$

Then, the system of equations is written as

$$\begin{aligned} B_{11}y_{1i} + B_{12}y_{2i} + C_1z_i &= u_{1i} \\ B_{21}y_{1i} + B_{22}y_{2i} + C_2z_i &= 0. \end{aligned} \tag{1.11}$$

$\underset{G_1}{B_{21}y_{1i}} + \underset{G-G_1}{B_{22}y_{2i}} + C_2z_i = 0.$

Now, If  $|B_{22}| \neq 0$ , we can solve the second set of equations for  $y_{2i}$ . That is, we have that

$$y_{2i} = -B_{22}^{-1}B_{21}y_{1i} - B_{22}^{-1}C_2z_i$$

and then the first set of equations in (1.11) becomes

$$\tilde{B}y_{1i} + \tilde{C}z_i = u_{1i}; \text{ or } \tilde{A}x_{1i} = u_{1i},$$

where

$$\begin{aligned}\tilde{B} &= B_{11} - B_{12}B_{22}^{-1}B_{21} \\ \tilde{C} &= C_1 - B_{12}B_{22}^{-1}C_2.\end{aligned}$$

If  $A_2$  were known, the previous model is nothing different than that from (1.1). But, in general is not the case.

The type of models we have seen so far are linear models. In practice, as in regression models, we can have that the parameters enter into the model in a nonlinear form.

## 1.2 NONLINEAR SEM

Again, we have a set of variables  $x_i = (y'_i, z'_i)'$ . Then, the relationship among the variables  $x_i$  is such that

$$u(x_i; \theta) = u_i$$

where  $u_i$  is the disturbance term, and  $u(\cdot, \cdot)$  is a possibly (known) nonlinear function in the set of parameters  $\theta_{(p \times 1)}$ . Again, about the disturbance term  $u_i$ , we assume either (1.2), (1.3) or (1.4) in addition to  $Eu_i = 0$ .

**Example 1** (a) *Linear Simultaneous Equation models*

$$u(x_i; \theta) = A(\theta)x_i.$$

(b) *Linear in  $y_i$ 's but not in  $z_i$*

$$By_i + H(z_i; \theta) = u_i$$

*The latter model is not of great interest, as it is a “quite” straightforward version of the linear one.*

The interesting models come when the nonlinearity is through the endogenous variables.

We already know some nonlinear models of this type.

**Example 2** *Box-Cox models where  $M = G = 1$ . Here*

$$u(x; \theta) = \begin{cases} \frac{y^\lambda - 1}{\lambda} - \beta' z & \lambda \neq 0 \\ \log y - \beta' z & \lambda = 0. \end{cases}$$

*Two models of interest are when*

$$\begin{array}{ll} \lambda = 0 & \lambda = 1 \\ \text{log-linear model} & \text{linear model} \end{array}$$

(i) *If we allowed  $\lambda$  unknown, then we might test whether the model is, say, log-linear or linear.*

(ii) *One implication of  $\lambda$  unknown is that the errors  $u_i$  cannot be Gaussian. The reason is because  $y_i \geq 0$ , and one further consequence for the Nonlinear Least Squares Estimator (NLLS) is that they can be inconsistent.*

**Example 3** *Arcsinh model with  $M = G = 1$ .*

$$u(x_i; \theta) = \frac{\operatorname{arcsinh}(\lambda y_i)}{\lambda} - \beta' z_i,$$

where  $\theta = (\lambda, \beta')'$  and

$$\operatorname{arcsinh}(z) = \ln \left( z + \sqrt{z^2 + 1} \right).$$

*This transformation does not suffer from the same problems as the Box-Cox. If  $\lambda = 0$  we then obtain the linear model.*

**Example 4**

As in the linear case:

$$\begin{aligned} u(x_i; \theta) &= u_i && \text{structural form equation} \\ \theta &&& \text{structural form parameter} \\ u_i &&& \text{structural form error} \\ \Sigma = E(u_i u_i') &&& \text{structural form covariance matrix.} \end{aligned}$$

However, what is the *REDUCED FORM*?

This is not as trivial as it was for the linear case. Suppose that in a neighbourhood of the true parameter vector  $\theta_0$ , we have that

$$u(x; \theta) = u \quad u(y, z; \theta) - u = 0$$

has a unique solution in  $y$ . That is,

$$y = R(u; z; \theta)$$

or equivalently

$$y_i = R(u_i, z_i, \theta), \quad i = 1, \dots, n.$$

### Example (Cont.)

1. The Box-Cox Model has a Reduced Form given by

$$y_i = \{1 + \lambda (\beta' z_i + u_i)\}^{\frac{1}{\lambda}}.$$

2. The arcsinh Model,

$$y_i = \frac{1}{\lambda} \sinh \{\lambda (\beta' z_i + u_i)\}.$$

3. Not explicit solution for  $R(\cdot)$ , but for given values of  $z$  and  $\theta$ , we can approximate the reduced form numerically.

## 2 IDENTIFICATION

### 2.1 BASIC DEFINITIONS AND EXAMPLES

In parametric inference, the identification of the unknown parameters  $\theta_{p \times 1}$  is defined through an *objective function*,

$$Q(x; \theta) = Q(\theta).$$

The function  $Q(\theta)$  measures the risk or loss incurred by the decision to adopt the value  $\theta$ . Then, we estimate an unknown true value  $\theta_0$  by incurring the minimum loss, i.e.

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta),$$

where  $\Theta \subset \mathbb{R}^p$  is a parameter space to be specified later.

**Example 5** Consider the linear regression model

$$y_i = \beta^t z_i + v_i, \quad i = 1, \dots, n.$$

If our objective function  $Q(\beta)$  is

$$Q(\beta) = (Y - Z\beta)'(Y - Z\beta),$$

then

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} Q(\beta),$$

which is the least squares estimator (LSE) of  $\beta$ . On the other hand, if we choose as our objective function

$$Q(\beta) = (Y - Z\beta)' \Sigma^{-1} (Y - Z\beta),$$

we then obtain the generalized least squares estimator

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} Q(\beta)$$

when  $\Sigma = E(vv')$ . If we knew the probability density function of  $v_i$ , then our objective function would become

$$Q(\theta) = -\log p(Y - \beta Z; \sigma^2),$$

where  $\theta = (\beta', \sigma^2)' \in \mathbb{R}^p \times (0, \infty)$ , and then you would be able to compute the Maximum Likelihood Estimator (MLE). The Instrumental variable estimator (IVE) is obtained as

$$\begin{aligned} Q(\beta) &= (Y - Z\beta)'(WW')(Y - Z\beta) \\ \tilde{\beta}^{IV} &= \arg \min_{\beta \in \mathbb{R}^p} Q(\beta) \\ &= (W'Z)^{-1} W'Y. \end{aligned}$$

**Example 6** *With nonlinear models, it is the same. If*

$$\begin{aligned} u(x_i; \theta) &= v(y_i; \lambda) - \beta' z_i \\ &= u_i, \end{aligned}$$

*e.g. Box-Cox or the arcsinh...If the probability density function of  $v_i$ ,  $p(v, \sigma^2)$ , were known then,*

$$Q(\theta) = - \sum_{i=1}^n \left\{ \log p(v(y_i; \lambda) - \beta' z_i; \sigma^2) - \underbrace{\log \left| \frac{\partial}{\partial y_i} v(y_i; \lambda) \right|}_{\text{Jacobian}} \right\}.$$

*In particular, if  $u_i \sim \mathcal{N}(0, \sigma^2)$*

$$\begin{aligned} Q(\theta) &= \frac{n}{2} \log 2\pi + \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} ((v(Y, \lambda) - Z\beta)' (v(Y, \lambda) - Z\beta)) \\ &\quad - \sum_{i=1}^n \log \left| \frac{\partial v(y_i, \lambda)}{\partial y_i} \right|. \end{aligned}$$

The pair  $(Q, \Theta)$  is thought of as containing all the information that will be used to choose a  $\theta$ -value. For a given data set  $X$ , two or more values give rise to the same  $Q$ , in which case one cannot choose between them by means of  $Q$ . Such an  $X$  may occur with small or zero probability. But if it exists for all possible  $X$ , there is some structural problem.

**Definition 4** *We say that two values  $\theta_1$  and  $\theta_2 \in \Theta$  are observationally equivalent (O.E.) with respect to  $Q$ , if*

$$Q(x; \theta_1) = Q(x; \theta_2) \quad \forall x \in \mathcal{X}.$$

**Example 7**

$$Q(\beta) = Y'Y + (\beta'Z' - 2Y')Z\beta$$

and  $\text{rank}(Z) < K$ , that is there is multicollinearity. Hence, because there exists  $c \neq 0$ , such that  $Zc = 0$ , we have that

$$\begin{aligned} Q(\beta) &= Y'Y + (\beta'Z' - 2Y')Z\beta \\ &= Y'Y + ((\beta + c)'Z' - 2Y')Z(\beta + c) \\ &= Q(\beta + c). \end{aligned}$$

Then, by definition

$\beta$  and  $\beta + c$  are O.E..

This implies that the LSE is not uniquely defined.

**Example 8**

$v_i \sim \mathcal{N}(0, \sigma^2)$  and  $\text{rank}(Z) < K \Rightarrow Zc = 0$  and  $c \neq 0$ .

Then,

$$\begin{aligned} Q \begin{pmatrix} \beta \\ \lambda \\ \sigma^2 \end{pmatrix} &= \frac{n}{2} \log 2\pi + \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} (v(Y, \lambda) - Z\beta)' (v(Y, \lambda) - Z\beta) \\ &\quad - \sum_{i=1}^n \log \left| \frac{\partial v(y_i; \lambda)}{\partial y_i} \right| \\ &= \frac{n}{2} \log 2\pi + \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} (v(Y, \lambda) - Z(\beta + c))' (v(Y, \lambda) - Z(\beta + c)) \\ &\quad - \sum_{i=1}^n \log \left| \frac{\partial v(y_i; \lambda)}{\partial y_i} \right| \\ &= Q \begin{pmatrix} \beta + c \\ \lambda \\ \sigma^2 \end{pmatrix}. \end{aligned}$$

*So, we conclude that*

$$\begin{pmatrix} \beta \\ \lambda \\ \sigma^2 \end{pmatrix} \text{ and } \begin{pmatrix} \beta + c \\ \lambda \\ \sigma^2 \end{pmatrix}$$

*are O.E.*

**Example 9** Let  $X$ ,  $Y$  and  $Z$  be

$$X = \underset{n \times (G+K)}{(x_1, \dots, x_n)'}; \quad Y = \underset{n \times G}{(y_1, \dots, y_n)'}; \quad Z = \underset{n \times K}{(z_1, \dots, z_n)'}$$

The observation matrix, e.g.  $X = (Y, Z)$ , is the matrix of observations. As usual the structural form parameters are  $(A, \Sigma)$ , whereas the reduced form parameters are  $(\Pi, \Omega)$ , where

$$\Pi = -B^{-1}C \quad \text{and} \quad \Omega = B^{-1}\Sigma(B^{-1})'$$

Assume that

$$\begin{aligned} v_i &\sim \text{Gaussian} \\ \Theta &= \{A, \Sigma : |B| \neq 0 \text{ and } \Sigma > 0\}. \end{aligned}$$

Now, because  $v_i \sim \text{Gaussian}$ , then

$$y_i | z_i \sim N(\Pi z_i; \Omega).$$

Therefore, our objective function  $Q$  is

$$Q(\theta) = Q(A, \Sigma) = \frac{1}{2}n \log 2\pi + \frac{1}{2}n \log |\Omega| \\ + \frac{1}{2} \operatorname{tr} \left\{ (Y - Z\Pi') \Omega^{-1} (Y - Z\Pi')' \right\},$$

*e.g.* it depends on  $A$  and  $\Sigma$  via  $\Pi$  and  $\Omega$ .

We know that knowledge of the likelihood function is equivalent to knowledge of  $\Pi$  and  $\Omega$ . Now, let  $P$  be a matrix such that  $|P| \neq 0$ , and

$$\bar{A} = PA; \quad \bar{\Sigma} = P\Sigma P'.$$

Then,

$$\begin{aligned} \bar{\Pi} &= -\bar{B}^{-1}\bar{C} \\ &= -(B^{-1}P^{-1})(PC) \\ &= -B^{-1}C \\ &= \Pi, \end{aligned}$$

and

$$\begin{aligned}
 \bar{\Omega} &= \bar{B}^{-1}\bar{\Sigma}(\bar{B}^{-1})' = (B^{-1}P^{-1})(P\Sigma P')(B^{-1}P^{-1})' \\
 &= B^{-1}\Sigma(B^{-1})' \\
 &= \Omega.
 \end{aligned}$$

Hence, we conclude that

$$Q(A, \Sigma) = Q(\bar{A}, \bar{\Sigma})$$

which implies that  $(A, \Sigma)$  and  $(\bar{A}, \bar{\Sigma})$  are O.E. by definition.

Example 9 illustrates that we can only deduce  $(A, \Sigma)$  from  $(\Pi, \Omega)$ . Moreover, the last two examples show that we cannot hope to estimate a regression model with multicollinearity or in a *Simultaneous Equation System*  $(A, \Sigma)$  with the only constraint that  $|B| \neq 0$ .

**Definition 5**  $\theta_1 \in \Theta$  is identifiable with respect to  $(Q, \Theta)$  if there is not any other  $\theta_2 \in \Theta$  such that  $\theta_2 \neq \theta_1$  which is O.E. to  $\theta_1$  with respect to  $Q$ . Otherwise,  $\theta_1$  is unidentifiable.

**Theorem 1** (*Sufficient condition for identifiability.*) *Let us assume that there exists a function  $\phi(x)$  such that  $\forall \theta \in \Theta$ , we have that*

$$\theta = \int \phi(x) f(Q(x; \theta)) dx, \quad (2.1)$$

*where  $f(\cdot)$  is some given function. Then,  $\theta$  is identifiable with respect to  $(Q; \theta)$ .*

**Proof.** Suppose the contrary. Then, we have that  $\forall x \in \mathcal{X}$ ,  $Q(x; \theta_1) = Q(x; \theta_2)$ . The latter implies that

$$\phi(x) f(Q(x; \theta_1)) = \phi(x) f(Q(x; \theta_2)),$$

and thus by (2.1) we obtain that

$$\begin{aligned} \theta_1 &= \int \phi(x) f(Q(x; \theta_1)) dx \\ &= \int \phi(x) f(Q(x; \theta_2)) dx \quad (\text{By the last displayed equality}) \\ &= \theta_2 \quad (\text{by (2.1)}). \end{aligned}$$

Therefore,  $\theta_1 = \theta_2$ , which concludes the proof.

■

**Example 10** (*Reduced Form*) Let

$$\Theta = \{\Pi, \Omega, \Omega > 0\} \quad \theta = \text{vec}[\Pi, \Omega],$$

and assume that there is no multicollinearity. Choose as our  $\phi(x)$  the function

$$\phi(x) = \text{vec}(\widehat{\Pi}, \widehat{\Omega}),$$

where

$$\widehat{\Pi} = Y'Z(Z'Z)^{-1} \quad \widehat{\Omega} = \frac{1}{n-K} Y' \left( I_n - Z(Z'Z)^{-1} Z' \right) Y.$$

Now, choose as the objective function the log-likelihood, and as  $f$

$$f = e^{-Q}.$$

Then, we obtain that

$$\int \phi(x) f(Q(x; \theta)) dx = E_{\theta} \left( \text{vec}(\widehat{\Pi}, \widehat{\Omega}) \right) = \text{vec} \left[ E_{\theta}(\widehat{\Pi}, \widehat{\Omega}) \right] = \theta.$$

*So, the parameters of the reduced form are identifiable if the model is not multicollinear.*

The next issue to examine is when the parameters of the structural form are identifiable.

## 2.2 IDENTIFICATION OF LINEAR SEM

This section focuses on the linear restrictions on  $A$  only.

We have that

$$\Pi = -B^{-1}C, \quad \text{that is,} \quad B\Pi + \underset{G \times K}{C} = 0$$

or equivalently

$$\underset{K \times G}{\Pi'} B' + C' = 0. \quad (2.2)$$

Thus, we have  $GK$  equations, but we have  $G^2 + GK$  unknowns. Then

$$\begin{aligned} \text{vec}(\Pi' B' + C') &= \text{vec}(\Pi' B' I_G) + \text{vec}(C') \\ &= (I_G \otimes \Pi') \text{vec}(B') + \text{vec}(C') \\ &= ((I_G \otimes \Pi'); I_{GK}) \underset{G(G+K) \times 1}{\alpha} \end{aligned} \quad (2.3)$$

where

$$\alpha = (\beta', \gamma')', \quad \beta = \text{vec}(B'), \quad \text{and} \quad \gamma = \text{vec}(C').$$

From here, we see clearly that we have

$$\begin{array}{ll} G(G+K) & \text{unknowns} \\ GK & \text{equations.} \end{array}$$

The latter implies that we need at least  $G(G+K) - GK$  extra equations to be able to identify  $A$ . Let  $W_{r \times G(G+K)}$  be a matrix (known) and  $W\alpha = w$  vector. Then, a necessary condition to identify  $A$  is that  $r \geq G^2$ . With this new set of extra constraints on the parameters  $\alpha$ , the system of equations becomes

$$\begin{aligned} [(I_G \otimes \Pi'); I_{GK}] \alpha &= 0 \\ W\alpha &= w, \end{aligned}$$

or in matrix notation

$$\begin{pmatrix} V \\ W \end{pmatrix} \alpha = \Psi \alpha = \begin{pmatrix} 0 \\ w \end{pmatrix} \neq 0.$$

**Theorem 2**  $\alpha$  is identified iff  $\text{rank}(\Psi) = G(G+K)$  (rank condition).

The condition  $r \geq G^2$  is known as the order condition.

Let  $D$  denote  $(I_G \otimes B; I_G \otimes C)$ .

**Theorem 3**  $\alpha$  identifiable iff  $\text{rank}(WD') = G^2$ .

**Proof.**

$$\begin{aligned} \Psi &= \begin{bmatrix} I_G \otimes \Pi' & I_{GK} \\ & W \end{bmatrix} = \begin{bmatrix} I_G \otimes \Pi' & I_{GK} \\ W_1 & W_2 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} 0_{G^2} & I_{GK} \\ WD' & W_2 \end{bmatrix}}_{\Psi_1} \underbrace{\begin{bmatrix} I_G \otimes (B')^{-1} & 0 \\ I_G \otimes \Pi' & I_{GK} \end{bmatrix}}_{\Psi_2}. \end{aligned}$$

Because  $\text{rank}(B) = G$ , it implies that  $\text{rank}(\Psi_2) = G(G + K)$ , and hence  $\text{rank}(\Psi) = \text{rank}(\Psi_1)$ .

However  $I_{GK}$  is a full rank matrix, so

$$\begin{aligned} \text{rank}(\Psi_1) &= GK + \text{rank}(WD') \\ &= G(G + K) \end{aligned}$$

if and only if

$$G^2 = \text{rank}(WD').$$

This concludes the proof. ■

**Remark 1** *When parameters are subject only to linear constraints (not only in linear SEM) we could equivalently reparametrize in terms of a set of unrestricted parameters.*

- Normally, the type of constraints that we have among the parameters are *homogeneous*, that is, some linear combination of the parameters is zero. For example, the exclusion restriction is homogeneous since the corresponding row of  $W$  and  $w$  are given, respectively, by

$$(0, 0, \dots, 0, 1, 0, \dots, 0); \quad \text{and} \quad 0.$$

Another example is

$$(0, \dots, 0, -1, 0, \dots, 0, 1, 0, \dots, 0); \quad \text{and} \quad 0.$$

- We always need at least one element of  $w$  to be different than zero, such as the normalization restriction where the corresponding element in  $w$  is 1.
- Typically we use the normalization

$$\text{diag}(B) = (1, 1, \dots, 1).$$

### 2.2.1 IDENTIFICATION OF A SINGLE EQUATION

Denote by  $\theta$  the underlying parameters of the system, that is

$$\theta = \begin{pmatrix} \theta'_1 & \theta'_2 \\ 1 \times p_1 & 1 \times p_2 \end{pmatrix}'; \quad \theta_0 = (\theta'_{01}; \theta'_{02}).$$

**Definition 6**  $\theta_{01}$  is identifiable w.r.t.  $(Q, \Theta)$  iff all  $\theta \in \Theta$  that are O.E. to  $\theta_0$  have the same value of  $\theta_1$ .

Let  $\alpha'_1 = (\beta'_1; \gamma'_1)$  be the parameters of the 1<sup>st</sup> equation, e.g. the 1<sup>st</sup> row of  $A$ . Then, in this case, what we have is that

$$\beta'_1 \Pi + \gamma'_1 = 0.$$

In this case, we have  $K$  equations and  $G + K$  unknowns, so that we need at least  $G$  extra equations to be able to identify  $\alpha_1$ .

Let  $r_1$  be additional constraints on  $\alpha_1$ ,

$$W_1 \alpha_1 = w_1.$$

**Theorem 4** (*Rank condition*) *The parameters*

$$\alpha_1 = \begin{pmatrix} \beta_1 \\ \gamma_1 \end{pmatrix}$$

*are identified iff*

$$\text{rank}(W_1 A') = G.$$

*( $r_1 \geq G$  order condition).*

**Proof.**

$$\left. \begin{array}{l} \Pi' \beta_1 + \gamma_1 = 0 \\ W_1 \alpha_1 = w_1 \end{array} \right\} = \begin{bmatrix} \Pi' & I_K \\ W_{11} & W_{12} \end{bmatrix} \alpha_1 = \begin{pmatrix} 0 \\ w_1 \end{pmatrix}.$$

Now,

$$\begin{pmatrix} \Pi' & I_K \\ W_{11} & W_{12} \\ r_1 \times G & r_1 \times K \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & I_K \\ W_1 A' & W_{12} \end{pmatrix}}_{\Psi_1} \times \underbrace{\begin{pmatrix} (B')^{-1} & 0 \\ \Pi' & I_K \end{pmatrix}}_{\Psi_2}.$$

Because  $|B| \neq 0$ , we have that  $\Psi_2$  is a full rank matrix, and then

$$\text{rank} \begin{bmatrix} \Pi' & I_K \\ W_{11} & W_{12} \end{bmatrix} = \text{rank} \begin{pmatrix} 0 & I_K \\ W_1 A' & W_{12} \end{pmatrix}.$$

From here we conclude that the the left side of the last displayed equality is  $G+K$  if and only if

$$\text{rank}(\Psi_1) = G + K = K + \text{rank}(W_1 A'),$$

which concludes the proof. ■

**Theorem 5** Assume that in the 1<sup>st</sup> equation all the constraints are zero, except the first one,  $\alpha_{11} = 1$ . Then  $\alpha_1$  is identifiable iff  $r_1 \geq G$  and  $\text{rank}(A^*) = G - 1$ , where  $A^*$  is the matrix formed from the second  $G - 1$  rows corresponding to zeroes in  $\alpha_1$ .

**Proof.** Rearrange  $x_i$  as  $\tilde{x}_i$ , and  $A$  as  $\tilde{A} = (\tilde{A}_1, \tilde{A}_2)$ , where all the restrictions are imposed in  $\tilde{A}_1 : G \times r_1$  and the normalization restriction is imposed on the  $(1, 1)$  element in  $\tilde{A}_1$ . Then,

$$W_1 = \left( I_{r_1} : 0 \right); \quad \tilde{A}'_1 = \left. \begin{array}{cc} 1 & \cdots \\ 0 & A^{*'} \end{array} \right\} r_1.$$

So,

$$W_1 \tilde{A}' = \tilde{A}'_1$$

and  $\text{rank}(W_1 \tilde{A}') = G$  if and only if  $\text{rank}(A^*) = G - 1$ . ■

Sometimes we have more constraints than we need for identification.

**Definition 7**  $\theta_0$  is overidentified if  $\exists$  two (or more) sets of prior constraints,

*each of which is capable of identifying  $\theta_0$ , and the union of the sets is linearly independent.*

$\theta_0$  is just-identified if it is identified but not overidentified. Important as far as the efficiency of estimators of  $\theta_0$  is concerned.

## 2.3 NONLINEARITY AND LOCAL IDENTIFICATION

**Example 11** *In a linear SEM, there could be constraints in the variance-covariance matrix of the structural form error  $u_i$ , i.e. such as*

$$\Sigma = \sigma^2 I \quad \text{or} \quad \Sigma = \lambda 11' + \sigma^2 I.$$

*(variance component models)*

*If we have restrictions*

$$w(A, \Sigma) = 0,$$

*then we can use*

$$B\Omega B' = \Sigma,$$

*as well as*

$$B\Pi + C = 0.$$

*But  $B\Omega B' = \Sigma$  is nonlinear in  $B$ .*

**Example 12** *If  $B$  is a lower triangular matrix with unit diagonal elements and  $\Sigma$  is a diagonal matrix, then the linear SEM is called a triangular SEM (or recursive system). E.g. Sims (1980)'s simple Macroeconomic model and its many variations afterward, or Cochrane (1994)'s permanent income model of consumption. The model can be equivalently represented as that where  $B$  is a lower triangular matrix with non-unit diagonals and  $\Sigma = I$ .*

**Example 13** Consider a simultaneous equation system with “AR” disturbances.

Then, the system can be written as

$$By_i + C^* z_i^* = u_i^* \quad |B| \neq 0$$

where  $z_i^*$  is a  $(K^* \times 1)$  vector

$$u_i^* = Du_{i-1}^* + e_i^*$$

with unknown  $D$  and

$$Ee_i^* = 0; \quad E(e_i e_j') = \Sigma \mathcal{I}(i = j).$$

Then

$$By_i + Cz_i = e_i$$

$$C = (C_1, C_2, C_3) \quad z_i = (y'_{i-1}, z_i^{*'}, z_{i-1}^{*'})'.$$

By definition, we know that

$$\left. \begin{array}{l} C_1 = -DB \\ C_2 = C^* \\ C_3 = -DC^* \end{array} \right\} \Rightarrow \left. \begin{array}{l} D = -C_1 B^{-1} \\ DC_2 + C_3 = 0 \end{array} \right\} \Rightarrow C_3 - C_1 B^{-1} C_2 = 0.$$

*Therefore, the last displayed expression induces nonlinear constraints on  $A$ . In particular there are  $GK^*$  of them. Thus, this set of constraints together with  $B\Pi + C = 0$  will imply that only  $G(G - K^*)$  extra constraints would be needed.*

**Example 14** *Another Example is to put restrictions on the impulse response function (dynamic causal effect) of a SVAR model. See for example Blanchard and Quah (1989) and Gali (1992). They typically concern the long-run cumulative effects of a structural shock  $j$  on another variable  $i$ . For instance, one may impose the hypothesis that the aggregate demand shocks do not have long-run effects on real GDP, or the long-run money neutrality. Given a SVAR and its Moving average representation (or impulse response function),*

$$A(L)y_t = u_t, \quad \Rightarrow \quad y_t = A(L)^{-1}u_t = D(L)u_t,$$

*the restriction takes the form*

$$D(1)_{ij} = 0.$$

*Note that*

$$D(1) = D_0 + D_1 + \dots,$$

*which is the reason why it is called “long-run restriction.” In the BQ’s original work,*

$$y_t = \begin{pmatrix} \Delta dgp_t \\ unempl_t \end{pmatrix} \text{ and } A(L) = B,$$

*and the long-run effect is imposed by setting  $B$  as a lower-triangular matrix. With added assumption of the diagonality of  $\Sigma$ , this model becomes the recursive model. With lagged term in  $A(L)$ , the identification is more complex.*

Suppose identifiability happens when there is a unique solution to a set of (possibly) nonlinear equations

$$\begin{matrix} \psi & (\theta) & = & 0, \\ q \times 1 & p \times 1 & & \end{matrix} \quad (2.4)$$

where  $\psi$  is a vector of given functions. Notice that in (2.4) we have suppressed any reference to known quantities.

We already know that if  $\psi(\cdot)$  is linear then there is either one unique solution to (2.4) or there are uncountable ones, (in fact, a continuous set of them). If  $\psi(\cdot)$  is nonlinear, then things are a bit different. We may have one, uncountable as before, solutions but in addition to these we may have also countable or finite number of solutions.

## FIGURES

**Definition 8**  $\theta_0$  is locally identifiable (L.I.) w.r.t.  $(Q, \theta)$  if  $\exists$  an open neighbourhood of  $\theta_0$  containing no other  $\theta$  which is O.E. to  $\theta_0$  w.r.t.  $Q$ .

**Definition 9**  $\theta_0$  is globally identifiable (G.I.) if and only if it is identifiable for every neighbourhood of  $\theta_0$ .

**Remark 2** (a) In the linear case, we have that G.I.  $\equiv$  L.I.. Indeed, if

$$W\alpha_1 = w \quad \text{and} \quad W\alpha_2 = w,$$

we then have that

$$W(\lambda\alpha_1 + (1 - \lambda)\alpha_2) = w$$

which it implies that

$$\lambda\alpha_1 + (1 - \lambda)\alpha_2 = \alpha_3$$

also satisfies the constraints. So, now let  $\lambda$  vary to conclude.

(b) If we can reduce  $\Theta$ , then L.I. becomes G.I., e.g. using inequality restrictions.

Let

$$\Psi(\theta) = \frac{\partial}{\partial \theta'} \psi(\theta), \quad \Psi_0 = \Psi(\theta_0).$$

**Definition 10**  $\theta_0$  is a regular point of  $\Psi(\theta)$  if  $\Psi(\theta)$  does not change its rank in a neighbourhood of  $\theta_0$ .

For an example, consider a restriction

$$\theta = (\theta_1, \theta_2)', \quad \psi(\theta) = \theta_1^2 + \theta_2^2 \quad (p = 2, q = 1).$$

Thus  $\theta = 0$  is identifiable, but

$$\text{rank}(\Psi(\theta)) = \text{rank} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{cases} 0 & \text{if } \theta = 0 \\ 1 & \text{if } \theta \neq 0 \end{cases},$$

Then  $\theta = 0$  is not a regular point of  $\Psi(\theta)$ .

**Theorem 6** *Let  $\theta_0$  be a solution of (2.4) ( $\psi(\theta) = 0$ ). Let  $\psi(\theta)$  be a continuous and differentiable function in a neighbourhood of  $\theta_0$ . Then,*

(a) *If  $\text{rank}(\Psi_0) = p$ , then  $\theta_0$  is L.I.*

(b) *If  $\theta_0$  is a regular point of  $\Psi(\theta)$  and  $\theta_0$  is L.I., then*

$$\text{rank}(\Psi_0) = p.$$

*We need obviously  $p \leq q$  for (a) and (b).*

**Proof.** We only prove part (a). First by differentiability of  $\psi(\theta)$ , the mean value theorem implies that

$$\psi(\theta) = \psi(\theta_0) + \Psi(\tilde{\theta})(\theta - \theta_0),$$

where  $\tilde{\theta}$  is an intermediate point between  $\theta$  and  $\theta_0$ . Suppose now that  $\theta_0$  is not L.I.. Then, by definition, there exists a sequence  $\theta^1, \theta^2, \dots, \theta^\ell, \dots$  converging to  $\theta_0$ , such that

$$\psi(\theta_0) = \psi(\theta^i) \quad i = 1, 2, \dots, \ell, \dots$$

Then, the last two expressions indicate that

$$\Psi(\tilde{\theta}^i) \frac{(\theta^i - \theta_0)}{|\theta^i - \theta_0|} = 0,$$

or that

$$\Psi(\tilde{\theta}^i) d^i = 0,$$

where  $\tilde{\theta}^i$  is an intermediate point between  $\theta^i$  and  $\theta_0$ . However, by definition  $d^i$ , is a vector lying in the unit circle and because  $\Psi(\cdot)$  is a continuous function, then

we obtain that in the limit

$$\Psi(\theta_0)d = 0.$$

So, the last equality implies that

$$\text{rank}(\Psi(\theta_0)) < p,$$

which concludes the proof of part (a). ■

Next we will study the case where there are constraints on  $\Sigma$ , e.g.

$$B\Pi + C = 0 \quad (1.16)$$

$$B\Omega B' = \Sigma \quad (1.17) \quad , \quad \text{and} \quad \theta = \text{vec}(B', C', \Sigma).$$

$$w(\theta) = 0 \quad (1.18)$$

Then we have the following theorem.

**Theorem 7** *Let  $\theta_0$  be a solution to (1.16) – (1.18). Let  $w(\theta)$  be a continuously differentiable function in a neighbourhood of  $\theta_0$ . Denote*

$$W(\theta) = \frac{\partial w(\theta)}{\partial \theta'}$$

$$H(\theta) = W(\theta) \begin{pmatrix} I_G \otimes B' \\ I_G \otimes C' \\ I_G \otimes 2\Sigma \end{pmatrix}.$$

Then,

- (a) *If  $\text{rank}(H(\theta_0)) = G^2$ , then  $\theta_0$  is L.I..*
- (b) *If  $\theta_0$  is a regular point of  $H(\cdot)$  and  $\theta_0$  is L.I., then  $\text{rank}(H(\theta_0)) = G^2$ .*

**Proof.** Let

$$\theta = (\beta', \gamma', \sigma')' = [\text{vec}'(B'); \text{vec}'(C'); \text{vec}(\Sigma)],$$

then

$$\Psi(\theta) = \begin{bmatrix} I_G \otimes \Pi' & I_{GK} & 0 \\ \Delta & 0 & -I_{GG} \\ W_\beta & W_\gamma & W_\sigma \end{bmatrix},$$

where

$$W_\beta = \left[ \frac{\partial w}{\partial \beta} \right]; \quad W_\gamma = \left[ \frac{\partial w}{\partial \gamma} \right]; \quad W_\sigma = \left[ \frac{\partial w}{\partial \sigma} \right].$$

But also  $\Psi(\theta)$  can be written as

$$\Psi(\theta_0) = \begin{pmatrix} 0 & I_{GK} & 0 \\ 0 & 0 & -I_{GG} \\ H^* & W_\gamma & W_\sigma \end{pmatrix} \begin{pmatrix} I_G \otimes (B')^{-1} & 0 & 0 \\ I_G \otimes \Pi' & I_{GK} & 0 \\ -\Delta & 0 & I_{GG} \end{pmatrix}, \quad (2.5)$$

where

$$H^* = W_\beta (I_G \otimes B') + W_\gamma (I_G \otimes C') + W_\sigma (I_G \otimes 2\Sigma).$$

cf.

$$\Delta = \frac{\partial \text{vec}(B\Omega B')}{\partial \beta'} = \frac{\partial}{\partial \beta'} (I_G \otimes B\Omega) \beta = 2 (I_G \otimes B\Omega) = 2 \left( I_G \otimes \Sigma (B')^{-1} \right)$$

So, because the rank of the second matrix on the right of (2.5) is full rank, then

$$\text{rank}(\Psi(\theta_0)) = \text{rank}(1^{\text{st}} \text{ matrix})$$

which is

$$G(G + K) + \text{rank}(H^*).$$

So, we apply Theorem 40 here to the matrix

$$H^* = W(\theta) \begin{pmatrix} I_G \otimes B' \\ I_G \otimes C' \\ I_G \otimes 2\Sigma \end{pmatrix}$$

to conclude the proof. ■

Global identification is difficult to establish in general. However, some restrictions may be helpful such as linearity, monotonicity. Revisit Example 12 and see Rubio-Ramírez et al (2010) as well.

## Identification of Transformed Parameters $\tau(\theta)$

Again our set of constraints are

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

Let

$$\begin{aligned}\Psi(\theta) &= \frac{\partial}{\partial \theta} \psi(\theta); & \Upsilon(\theta) &= \frac{\partial \tau(\theta)}{\partial \theta'} \\ \Psi_0 &= \Psi(\theta_0); & \Upsilon_0 &= \Upsilon(\theta_0).\end{aligned}$$

**Theorem 8** (a) *If*

$$\text{rank} \begin{pmatrix} \Psi_0 \\ \Upsilon_0 \end{pmatrix} = \text{rank}(\Psi_0),$$

*then  $\tau_0 = \tau(\theta_0)$  is L.I., e.g. there is not any other value of  $\theta$  in  $\mathcal{N}(\theta_0)$  such that*

$$\psi(\theta) = 0 \quad \text{and} \quad \tau(\theta) \neq \tau_0.$$

(b) If  $\theta_0$  is a regular point of  $\begin{pmatrix} \Psi(\theta) \\ \Upsilon(\theta) \end{pmatrix}$  and  $\Upsilon(\theta)$  and  $\tau_0 = \tau(\theta_0)$  is L.I.,  
then

$$\text{rank} \begin{pmatrix} \Psi_0 \\ \Upsilon_0 \end{pmatrix} = \text{rank}(\Psi_0).$$

**Proof.** The proof is heuristic. The L.I. that

$$\psi(\theta) = 0 \text{ implies that } \tau(\theta) = \tau_0,$$

and the differentiability of  $\tau$  and  $\psi$  implies that there exists a transformation  $g$  such that

$$\tau(\theta) = g(\psi(\theta)), \quad \forall \theta \in \mathcal{N}(\theta_0).$$

Then,

$$\begin{aligned} \Upsilon(\theta) &= \frac{\partial \tau(\theta)}{\partial \theta} = \frac{\partial g}{\partial \Psi'} \frac{\partial \psi}{\partial \theta'} \\ &= G(\theta) \Psi(\theta). \end{aligned}$$

So that

$$\begin{aligned} \text{rank} \begin{pmatrix} \Psi_0 \\ \Upsilon_0 \end{pmatrix} &= \text{rank} \begin{pmatrix} \Psi_0 \\ G_0 \Psi_0 \end{pmatrix} \\ &= \text{rank} \left[ \begin{pmatrix} I \\ G_0 \end{pmatrix} \Psi_0 \right] = \text{rank}(\Psi_0). \end{aligned}$$

This concludes the proof. ■

**Example 15** Let  $\tau(\theta) = \theta$ . Then

$$\text{rank} \begin{pmatrix} \Psi_0 \\ I_p \end{pmatrix} = \text{rank}(\Psi_0).$$

iff  $\text{rank}(\Psi_0) = p$ . (We are implicitly assuming that  $\theta_0$  is a regular point.) Thus,  $\tau_0$  is L.I. iff  $\theta_0$  is L.I.

**Example 16** Suppose

$$\psi(\theta) = \begin{pmatrix} \theta_1 - \theta_2 \\ \theta_3 - 1 \end{pmatrix} \quad \text{and} \quad \tau(\theta) = \theta_1 - \theta_2.$$

*Then,*

$$\Psi = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \Upsilon = (1 \ -1 \ 0),$$

*which satisfies the condition of the above theorem.*

## 2.4 IDENTIFIABILITY AND ASYMPTOTIC THEORY

Let the *Objective Function* be

$$Q_n(X; \theta)$$

such that

$$p \lim Q_n(X; \theta) = \bar{Q}(\theta).$$

uniformly in  $\theta \in \Theta$ . Let “true value”  $\theta_0$  satisfy that

$$\theta_0 = \arg \min_{\theta \in \Theta} \bar{Q}(\theta). \tag{2.6}$$

This is very reasonable because it makes sense to estimate  $\theta_0$  by

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q_n(X; \theta).$$

Let

$$R(\theta) = \bar{Q}(\theta) - \bar{Q}(\theta_0).$$

Because  $\theta_0$  satisfies (2.6), then

$$R(\theta) \geq 0 \quad \text{and} \quad R(\theta_0) = 0.$$

Consider the following conditions:

$$(a) \quad R(\theta) > 0 \quad \forall \theta \in \Theta - \{\theta_0\}.$$

$$(b) \quad R(\theta) > 0 \quad \forall \theta \in \mathcal{N}(\theta_0) - \{\theta_0\}.$$

**Theorem 9** *If condition (a) holds then  $\theta_0$  is G.I., whereas if condition (b) holds then  $\theta_0$  is L.I..*

**Proof.** Suppose that  $\theta_1, \theta_0$  are O.E., then by definition we have that  $Q_n(x, \theta_1) = Q_n(x, \theta_0), \forall x \in \mathcal{X}$ , and hence

$$\bar{Q}(\theta_1) = \bar{Q}(\theta_0)$$

which implies that  $R(\theta_1) = 0$ . But  $R(\theta) > 0$  for all  $\theta \in \Theta - \{\theta_0\}$ . Then, we conclude that  $\theta_1, \theta_2$  are not O.E., or that  $\theta_0$  is *G.I.*. The proof for *L.I.* is identical and thus it is omitted. ■

## FIGURES

**Remark 3** *Conditions (a) and (b) are sufficient but not necessary.*

**Definition 11** *If condition (a) holds then we say that  $\theta_0$  is asymptotically identifiable (A.I.). On the other hand, if condition (b) holds then we say that  $\theta_0$  is asymptotically locally identifiable (A.L.I.).*

In fact, this is all we need for the consistency of the estimators.

**Example 17** Consider the linear regression model

$$y_i = \beta' z_i + v_i,$$

where we have that

$$\widehat{M} = \frac{Z'Z}{n} \xrightarrow{P} M.$$

Then,

- (a) If  $\widehat{M} > 0$ , then  $\beta$  is identifiable.
- (b) If  $M > 0$ , we then have that  $\beta$  is A.I..

**Theorem 10** Let  $\bar{Q}(\theta)$  be twice continuously differentiable in a neighbourhood of  $\theta_0$  and put

$$S(\theta) = \frac{\partial^2 \bar{Q}(\theta)}{\partial \theta \partial \theta'}.$$

Then

$$\text{rank}(S(\theta_0)) = p \Rightarrow \theta_0 \text{ is A.L.I.}$$

If  $\theta_0$  is a regular point of  $S(\theta)$ , and  $\theta_0$  is A.L.I., then

$$\text{rank}(S(\theta_0)) = p.$$

**Remark 4** The matrix  $S^{-1}(\theta)$  arises in the Central Limit Theorem for estimates minimizing  $Q_n(x; \theta)$ . Thus A.L.I. is closely related to conditions for asymptotic normality, not to mention the consistency.

### 3 ESTIMATION OF LINEAR SEM

#### 3.1 PARAMETERIZATIONS

Consider the linear simultaneous equation model

$$Ax_i = u_i; \quad x_i = \begin{pmatrix} y_i \\ z_i \end{pmatrix}_{(G+K) \times 1},$$

where the system is complete and  $|B| \neq 0$ . Furthermore, the variables  $z_i$  are exogenous and

1.  $E(u_i | z_i) = E(u_i) = 0$
2.  $u_i$  is homoscedastic and serially uncorrelated, that is

$$\begin{aligned} E(u_i u_i' | z_i) &= E(u_i u_i') = \Sigma_0 > 0 \\ E(u_i u_j' | \{z_i\}_{i=1}^n) &= E(u_i u_j') = \Sigma_0 \mathcal{I}(i = j). \end{aligned}$$

Possible parameterizations of interest are

(a) Structural Form

$$(A, \Sigma) = \Phi$$

identified only if there are constraints. Nevertheless,  $A$  should be economically sounded.

(b) Reduced Form

$$(\Pi, \Omega) = \Psi,$$

which is always identified unless  $z_i$  is collinear. Possible interest on  $\Psi$  is for prediction purposes of  $Y$ . That is,  $\hat{y}_i = \hat{\Pi}z_i$ , which is the *LSE* predictor of  $Y|Z$ , where  $\hat{\Pi}$  is the *LSE* of  $\Pi_0$ . Properties of  $\hat{y}_i$  depends on those of  $\hat{\Pi}$ .

(c) Minimal Parameters

Let  $\theta_{p \times 1}$  be a set of parameters ( $p \leq GK + \frac{1}{2}G(G+1)$ ) of functionally unrelated parameters.

Recall that if  $p = GK + \frac{1}{2}G(G + 1)$ , then the model is *just-identified*. So,

$$\begin{aligned}\Phi &= \Phi(\theta) = (A(\theta), \Sigma(\theta)) \\ \Psi &= \Psi(\theta) = (\Pi(\theta), \Omega(\theta)).\end{aligned}$$

If  $p < GK + \frac{1}{2}G(G + 1)$ , then the model is *over-identified*.

Thus, basically what we are saying is that the matrix  $(A, \Sigma)$  depends on a set of parameters  $\theta$  which we are concerned with. If the set of constraints are linear, then we can give an explicit solution for  $\Phi(\theta)$ , that is

$$\left. \begin{aligned}B\Pi + C &= 0 \\ W\alpha &= w.\end{aligned} \right\}$$

When there are nonlinear constraints, we can also do the same, though  $A(\theta)$  will be only *Implicitly* defined. Thus, only local approximations of  $A(\theta)$  around  $\theta$  can be made and corresponding numerical values of  $\Phi, \theta$  can be determined.

In many situations, the parameters involving “A” and “ $\Sigma$ ” can be independent

$$(A(\theta), \Sigma(\theta)) = (A(\theta_1), \Sigma(\theta_2)), \quad \theta = (\theta'_1, \theta'_2)'$$

Many times only “ $A$ ” is of interest, thus we may consider only  $A = A(\theta_1)$ , where now  $p \leq GK$ .

- The question is, how do we estimate  $\theta$ ?
  1. MLE principle: PMLE
  2. Least Squares Principle:
    - (a) MDE for RF error
    - (b) 2SLS and 3SLS for SF error

## 3.2 GAUSSIAN PSEUDO-MAXIMUM LIKELIHOOD ESTIMATOR (PMLE)

Some authors call this estimator *QMLE* (*Quasi-Maximum Likelihood Estimator*).

Consider the  $-\log$  times  $2/n$  as an objective function, that is

$$\begin{aligned} Q(\theta) &= C + \log |\Omega| + \frac{1}{n} \text{tr} \left[ (Y - Z\Pi')' (Y - Z\Pi') \Omega^{-1} \right] \\ &= C + \log |\Sigma| - 2 \log |B| + \text{tr} \left[ A \frac{X'X}{n} A' \Sigma^{-1} \right] \\ &= Q_1(\Phi) \\ &= Q_2(\Psi), \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} \Phi &= \Phi(\theta), \quad \Psi = \Psi(\theta), \quad \Pi = \Pi(\theta) \\ \Sigma &= \Sigma(\theta), \quad \Omega = \Omega(\theta), \quad A = A(\theta). \end{aligned}$$

**Definition 12** *Let*

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta).$$

*Then  $(\hat{\theta}, \hat{\Phi}, \hat{\Psi})$  is the (Gaussian) PMLE of  $(\theta_0, \Phi_0, \Psi_0)$  with  $\hat{\Phi} = \Phi(\hat{\theta})$  and  $\hat{\Psi} = \Psi(\hat{\theta})$ .*

If  $(u_i | z_i) \sim NID(0, \Sigma_0)$ , these estimates become the (full information) MLE or *FIML*.

Let's introduce the following assumptions.

**A1**  $\Theta$  is a compact set in  $\mathbb{R}^p$ , with  $|B| \neq 0$ ,  $\Sigma > 0$ ,  $\forall \theta \in \Theta$ .

**A2**  $\Phi$  is a continuous function in  $\theta$ .

A2 also implies that  $\Psi$  is also continuous in  $\theta$ .

**Theorem 11** *Under A1 and A2, the PMLE exists, although it may not be unique.*

**Proof.**  $Q(\cdot)$  is a continuous function of  $(A, \Sigma)$ , so that it is also in  $\theta$  by  $A2$ . But  $\theta \in \Theta$  is a compact set, so it implies that the minimum exists. ■

**Theorem 12** *If  $\Sigma$  is unconstrained and  $\theta$  only parameterizes  $A$ , then we have that the PMLE of  $\theta_0$  is given by*

$$\hat{\theta} = \arg \min_{\theta \in \Theta} R(\theta),$$

where

$$\begin{aligned} R(\theta) &= \left( \frac{|B|^2}{|A \frac{X'X}{n} A'|} \right)^{-1} \\ &= \left| \frac{(Y - Z\Pi)'(Y - Z\Pi)}{n} \right| = \left| \frac{V'V}{n} \right|, \end{aligned} \tag{3.2}$$

and the PMLE of  $\Sigma$  is

$$\hat{\Sigma} = \hat{A} \frac{X'X}{n} \hat{A}'.$$

**Proof.** By definition, the *PMLE* is  $\hat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta)$ . Now,

$$\begin{aligned}
Q_1(\Phi) &= \log |\Sigma| - 2 \log |B| + \frac{1}{n} \text{tr} (AX'XA'\Sigma^{-1}) \\
&= \log |AX'XA'| - \log |AX'XA'\Sigma^{-1}| \\
&\quad - 2 \log |B| + \frac{1}{n} \text{tr} (AX'XA'\Sigma^{-1}) \\
&= \log |AX'XA'| - \log \left\{ \left| \Sigma^{-\frac{1}{2}} A \frac{X'X}{n} A' \Sigma^{-\frac{1}{2}} \right| n^G \right\} \\
&\quad - 2 \log |B| + \text{tr} \left( \Sigma^{-\frac{1}{2}} A \frac{X'X}{n} A' \Sigma^{-\frac{1}{2}} \right).
\end{aligned}$$

But,

$$\begin{aligned}
\log |AX'XA'| - 2 \log |B| &= \log \left| B^{-1} AX'XA' (B')^{-1} \right| \\
&= \log \left| B^{-1} (B, C) \begin{pmatrix} Y' \\ Z' \end{pmatrix} (Y, Z) \begin{pmatrix} B' \\ C' \end{pmatrix} (B')^{-1} \right| \\
&= \log |(Y' - \Pi Z')(Y - Z\Pi')| \\
&= \log |R(\theta)| + G \log n.
\end{aligned}$$

Denote

$$D = \Sigma^{-\frac{1}{2}} A \frac{X'X}{n} A' \Sigma^{-\frac{1}{2}}.$$

Then

$$Q_1(\Phi) = \log |R(\theta)| - \log |D| + \text{tr}(D).$$

On the other hand, by definition,

$$\text{tr}(D) - \log |D| = \sum_{j=1}^G (\lambda_j - \log \lambda_j).$$

Thus

$$Q_1(\Phi) > \log |R(\theta)|$$

which implies that

$$Q_1(\Phi) = \begin{cases} \geq \log |R(\theta)| + G \\ = \log |R(\theta)| + G \text{ iff } \lambda_j = 1 \ \forall j, \end{cases}$$

e.g.

$$\Sigma^{-\frac{1}{2}} \frac{A X' X A}{n} \Sigma^{-\frac{1}{2}} = I_G.$$

Thus, we minimize  $Q_1(\Phi)$  by minimizing  $|R(\theta)|$  with respect to  $\theta$  and then,

$$\widehat{\Sigma} = \widehat{A} \frac{X'X}{n} \widehat{A}',$$

where recall  $\widehat{A} = A(\widehat{\theta})$ . ■

### 3.3 MINIMUM DISTANCE ESTIMATOR (MDE) (NLLS)

Another way to estimate  $\theta$ , which only parametrizes  $\Pi$ , is to plug in  $\hat{\Omega}$  from the unrestricted OLS. That is,

**Definition 13**

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta \in \Theta} \left( Q_2 \left( \Pi, \hat{\Omega} \right) - \log \left| \hat{\Omega} \right| \right) \\ \hat{A} &= A \left( \hat{\theta} \right); \quad \hat{\Pi} = \Pi \left( \hat{\theta} \right) \\ \hat{\Omega} &= \frac{1}{n} Y' \left( I - Z \left( Z' Z \right)^{-1} Z' \right) Y\end{aligned}$$

are the MDE of  $(\theta_0, A_0, \Pi_0)$  respectively.

It is easy to see that the MDE is identical to the one that minimizes

$$\frac{1}{n} \text{tr} \left\{ V' V \hat{\Omega}^{-1} \right\},$$

and the term ‘least squares’ makes sense.

Let’s introduce the following assumptions

**B1**  $\Theta \subset \mathbb{R}^p$  is a compact set such that  $|B| \neq 0$  for all  $\theta \in \Theta$ .

**B2**  $A$  is a continuous function in  $\theta$ .

**Theorem 13** *Assuming B1 and B2, the MDE exists. (Not need to be unique.)*

### 3.4 TWO STAGE AND THREE STAGE LEAST SQUARES

The previous two estimators, that is the *PMLE* and the *MDE*, involve numerical optimization (they do not have a close form, except in situations where the parameters are just-identified). The latter is true even when the model is linear in “ $A$ ” and “ $A$ ” is maybe linear in  $\theta$ . In the latter case, we can obtain estimates which have an explicit formula.

**Definition 14** For  $\widehat{\Sigma}$  defined below, we define the 3SLS estimator of  $(\theta_0, A_0, \Pi_0)$ , denoted  $(\widehat{\theta}, \widehat{A}, \widehat{\Pi})$ , as the value that minimizes the objective function

$$S_{3SLS}(\theta) = \frac{1}{n} \text{tr} \left\{ U' Z (Z' Z)^{-1} Z' U \widehat{\Sigma}^{-1} \right\}.$$

That is,

$$\widehat{\theta} = \arg \min_{\theta \in \Theta} S_{3SLS}(\theta)$$

and then

$$\widehat{A} = A(\widehat{\theta}); \quad \widehat{\Pi} = \Pi(\widehat{\theta}).$$

The 3SLS is then the two-step optimal GMM estimator under our maintained assumption. It can be motivated different ways. Indeed, consider the following objective function

$$Q_1(\Phi) = \log |B^{-1}\Sigma B'^{-1}| + \frac{1}{n} \text{tr} (AX'XA'\Sigma^{-1}).$$

Given  $\Sigma$ , the nonlinearity in the estimation enters through the term  $\log |B^{-1}\Sigma B'^{-1}|$ .

Now, recall that

$$Q_1(\Phi) = Q_2(\Pi, B^{-1}\Sigma B'^{-1}).$$

Next, consider the *LSE* of  $\Pi$ ,  $\tilde{\Pi} = Y'Z(Z'Z)^{-1}$ . It obvious that for given  $\Sigma$ , say  $\hat{\Sigma}$ ,

$$Q_2(\tilde{\Pi}, B^{-1}\hat{\Sigma}B'^{-1}) \leq Q_2(\Pi, B^{-1}\hat{\Sigma}B'^{-1})$$

for all  $\Pi$ . Then,

$$\begin{aligned}
S_{SLS}(\theta) &= Q_2\left(\Pi, B^{-1}\widehat{\Sigma}B'^{-1}\right) - Q_2\left(\widetilde{\Pi}, B^{-1}\widehat{\Sigma}B'^{-1}\right) \\
&= \frac{1}{n}tr\left\{\left(V'V - \widetilde{V}'\widetilde{V}\right)\left(B^{-1}\widehat{\Sigma}B'^{-1}\right)^{-1}\right\} \\
&= \frac{1}{n}tr\left\{\left(\widetilde{\Pi} - \Pi\right)Z'Z\left(\widetilde{\Pi} - \Pi\right)'\left(B^{-1}\widehat{\Sigma}B'^{-1}\right)^{-1}\right\} \\
&= \frac{1}{n}tr\left\{A\begin{pmatrix} \widetilde{\Pi} \\ I \end{pmatrix}Z'Z\begin{pmatrix} \widetilde{\Pi} \\ I \end{pmatrix}'A'\widehat{\Sigma}^{-1}\right\} \\
&= \frac{1}{n}tr\left\{AX'Z\left(Z'Z\right)^{-1}Z'XA'\widehat{\Sigma}^{-1}\right\},
\end{aligned}$$

as  $\widehat{V} + Z\left(\widehat{\Pi} - \Pi\right)' = V$  for the third equality.

**Theorem 14** *Under B1 and B2 the 3SLS exists. (Need not to be unique.)*

We have seen in the previous definition that for the definition of the 3SLS we need to estimate  $\Sigma$ . How? The next estimator will provide the answer to this

question. Define the following objective function

$$\begin{aligned} S_{2SLS}(\theta) &= Q_2(\Pi, B^{-1}B'^{-1}) - Q_2(\tilde{\Pi}, B^{-1}B'^{-1}) \\ &= \frac{1}{n} \text{tr} \left\{ AX'Z (Z'Z)^{-1} Z'XA' \right\}. \end{aligned}$$

Observe that it is the same objective function  $S_{3SLS}(\theta)$  but with  $\hat{\Sigma}$  being replaced by  $I$ , the identity matrix.

**Definition 15** We define the 2SLS estimator of  $(\theta_0, A_0, \Pi_0)$ , denoted  $(\tilde{\theta}, \tilde{A}, \tilde{\Pi})$ , as the value that minimizes the objective function

$$\begin{aligned} S_{2SLS}(\theta) &= \frac{1}{n} \text{tr} \left\{ A \begin{pmatrix} \tilde{\Pi} \\ I \end{pmatrix} Z'Z \begin{pmatrix} \tilde{\Pi} \\ I \end{pmatrix}' A' \right\} \\ &= \frac{1}{n} \text{tr} \left\{ AX'Z (Z'Z)^{-1} Z'XA' \right\}. \end{aligned}$$

That is,

$$\tilde{\theta} = \arg \min_{\theta \in \Theta} S_{2SLS}(\theta)$$

and then

$$\tilde{A} = A(\tilde{\theta}); \quad \tilde{\Pi} = \Pi(\tilde{\theta}).$$

**Theorem 15** *Under B1 and B2, the 2SLS exists, (need not to be unique), and*

$$\hat{\Sigma} = \tilde{A} \frac{X'X}{n} \tilde{A}'.$$

## Remarks

1.  $\text{tr}\{U'PU\} = \sum_{g=1}^G U'_g P U_g$ , where  $U = (U_1, \dots, U_G)$  and  $P = Z(Z'Z)^{-1}Z'$ .
2. If all the restrictions are linear, imposing them yield a linear regression for each equation  $g$

$$\tilde{y}_{gi} = \tilde{z}'_{gi} \theta_g + u_{gi}, \quad g = 1, \dots, G, \quad i = 1, \dots, n,$$

where  $\tilde{y}_{gi}$  and  $\tilde{z}_{gi}$  are linear transformations of  $x_i$ .

- Thus,  $\tilde{z}_{gi}$  may suffer from endogeneity.
- The dimension of  $\theta_g$  depends on the number of restrictions imposed on the  $g$ -th equation.
- If there is no cross-equation restriction, then the 2SLS is equivalent to equationwise 2SLS since there is no common element in  $\theta_g$  and  $\theta_f$  for  $f \neq g$ .

3. The 3SLS is the optimal GMM, that is,

$$\frac{1}{n} \text{tr} \left\{ U' Z (Z' Z)^{-1} Z' U \widehat{\Sigma}^{-1} \right\} = m_n(\theta)' W_n m_n(\theta)$$

where  $m_n(\theta) = \frac{1}{n} \sum_{i=1}^n u_i(\theta) \otimes z_i$  and  $W_n = n \left( \widehat{\Sigma} \otimes Z' Z \right)^{-1}$ . And if all the restrictions are linear,  $m_n(\theta)$  is linear in  $\theta$ , yielding a closed-form solution<sup>2</sup>.

---

<sup>2</sup>For  $g = 1, \dots, G$ ,  $u_{gi}(\theta) z_i = \tilde{y}_{gi} z_i - z_i \tilde{z}'_{gi} \theta_g$  and thus  $u_i(\theta) \otimes z_i = \tilde{y}_i - \tilde{\mathbf{Z}}_i \theta$ , where  $\tilde{y}_i = (\dots \tilde{y}_{gi} z'_i \dots)'$  and  $\tilde{\mathbf{Z}}_i$  is the matrix of  $z_i \tilde{z}'_{gi}$ s constructed conformably to  $\theta$ . Note that  $\theta_g, g = 1, \dots, G$  may contain some common elements due to the cross-equation restriction. Otherwise,  $\tilde{\mathbf{Z}}_i$  is a block-diagonal matrix.

### 3.5 INDIRECT LEAST SQUARES (ILSE). RELATIONSHIP BETWEEN DIFFERENT ESTIMATORS

The indirect least squares estimators (ILSE)  $\hat{\theta}$  and  $\hat{A}$  of  $\theta_0$  and  $A_0$  satisfy

$$\hat{A} \begin{pmatrix} \tilde{\Pi} \\ I \end{pmatrix} = 0, \text{ or } \hat{B}\tilde{\Pi} + \hat{C} = 0 \quad (5.4)$$

where  $\hat{A} = A(\hat{\theta})$ ,  $\tilde{\Pi}$  is the *LSE* of  $\Pi$ , and  $p \leq GK$ .

**Theorem 16** *Suppose that there are not overidentifiability constraints and there is a unique solution for the ILSE. Then,*

$$ILSE = PMLE = MDE = 3SLSE = 2SLSE.$$

**Proof.** Note that

$$\begin{aligned}
 \hat{\Pi}_{PMLE} &= \underset{\Pi}{\operatorname{argmin}} Q_2 \left( \Pi, \hat{\Omega}_{PMLE} \right) \\
 &= \underset{\Pi}{\operatorname{argmin}} \operatorname{tr} \left( V' V \hat{\Omega}_{PMLE}^{-1} \right) \\
 &= \underset{\Pi}{\operatorname{argmin}} \operatorname{tr} (V' V) \\
 &= \tilde{\Pi},
 \end{aligned}$$

and the same is true for *MDE*.

Regarding *3SLS*, it is obvious due to its loss function

$$S_{3SLS}(\theta) = Q_2 \left( \Pi, B^{-1} \hat{\Sigma} B'^{-1} \right) - Q_2 \left( \tilde{\Pi}, B^{-1} \hat{\Sigma} B'^{-1} \right)$$

and its minimum is achieved at  $\Pi = \tilde{\Pi}$  and thus  $\hat{\Pi}_{3SLS} = \tilde{\Pi}$ . ■

**Example 18** Consider the case where all the restrictions are linear

$$\begin{aligned} [(I_G \otimes \Pi') ; I_{GK}] \alpha &= 0 \\ W\alpha &= w, \end{aligned}$$

that is,

$$\Psi\alpha = \psi.$$

Then,

$$\hat{\alpha}_{ILS} = \hat{\Psi}^{-1}\psi,$$

where  $\hat{\Psi}$  is the plug-in estimator using the OLSE of  $\Pi$ .

To understand the equivalence between ILS, 3SLS, recall that imposing the linear restrictions yield a linear regression for each equation  $g$

$$\tilde{y}_{gi} = \tilde{z}'_{gi}\theta_g + u_{gi}, \quad g = 1, \dots, G, \quad i = 1, \dots, n,$$

where  $\tilde{y}_{gi}$  and  $\tilde{z}_{gi}$  are linear transformations of  $x_i$ . Write it in matrix notation

$$\dot{y}_g = \dot{Z}_g\theta_g + U_g,$$

where  $\theta'_g$ 's dimension is  $G$  as each equation is just-identified. Let

$$\ddot{Z}_g = P_Z \dot{Z}_g = Z Q_g = Z (Z' Z)^{-1} Z' \dot{Z}_g,$$

then  $Q_g$  is a p.d. square matrix. Recall that the 2SLS is the OLS of

$$\begin{aligned} \dot{y}_g &= \ddot{Z}_g \theta_g + U_g, \\ &= Z \pi_g + U_g, \end{aligned}$$

where

$$Q_g \theta_g = \pi_g.$$

Furthermore, for a  $R$ , again reflecting the linear restrictions, we can write

$$\begin{aligned} \dot{Y} &= (\dot{y}_1, \dots, \dot{y}_G) = X R = (Y : Z) \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} \\ &= Z (\pi_1, \dots, \pi_G) + U \\ &= Z \Pi' + U. \end{aligned}$$

*And*

$$\hat{\Pi}' = (Z'Z)^{-1} Z'\dot{Y} = (Z'Z)^{-1} Z'(YR_1 + ZR_2) = \tilde{\Pi}'_{OLS}R_1 + R_2.$$

*That is,  $\hat{\Pi}$  is a linear transformation of the OLSE of  $\Pi$  and thus,  $\hat{\theta}$  is. Thus, the 2SLS is the ILS. Next, the GLS is the OLS in the multiple regression (that of  $\dot{Y}$  on  $\ddot{Z}$ ). This leads us to conclude that the 3SLS is ILS.*

**Example 19** *Estimation of the recursive system is easy. The Cholesky decomposition of a given  $\Omega$  produces the corresponding  $B$  directly, that is, apply the decomposition to write*

$$\Omega = LL',$$

where  $L$  is known to be unique. Then we have

$$B = \Sigma^{\frac{1}{2}}L^{-1}.$$

As  $B$  has unit diagonals,  $\Sigma^{1/2}$  should be the diagonal matrix whose diagonals are the inverses of those of  $L^{-1}$ . Thus, let

$$\hat{\Omega} = \frac{1}{n}\hat{V}'\hat{V} = \hat{L}\hat{L}', \quad (1)$$

where  $\hat{U}$  is the matrix stacking the OLS residuals  $\hat{u}_t$ 's. Then,  $\hat{\Sigma}$  is the diagonal matrix of the inverse of the diagonals of  $\hat{L}^{-1}$  and

$$\hat{B} = \hat{\Sigma}^{\frac{1}{2}}\hat{L}^{-1}.$$

### 3.6 SINGLE EQUATION AND SUBSYSTEM ESTIMATION

We are now concerned with the estimation of one or several equations, say  $G_1$ , where  $1 \leq G_1 < G$ .

Why? Some possible reasons are

- 1 The whole system is not identified but one or more equations are.
- 2 Because those  $G_1$  equations are the ones we are interested in.
- 3 Easier to compute, although we may lose some efficiency.

Let's introduce the following notation.

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \quad \begin{array}{l} \rightarrow G_1 \times (G + K) \\ \rightarrow (G - G_1) \times (G + K). \end{array}$$

$$Ax_i = \begin{bmatrix} A_1x_i \\ A_2x_i \end{bmatrix} = \begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix};$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}; \quad A_1 = [B_{11}; B_{12}; C_1]$$

$$A_2 = [B_{21}; B_{22}; C_2].$$

Also, let

$$y_i = \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix};$$

with  $|B_{11}| \neq 0$ , and we shall assume that  $A_1$  and  $\Sigma_{11}$  are identifiable. It is worth mentioning that because we are concerned with the first  $G_1$  equations,  $A_2, \Sigma_{12}, \Sigma_{22}$  may not be identifiable.

Let's examine  $A_2x_i = u_{2i}$ . The latter expression is

$$B_{21}y_{1i} + B_{22}y_{2i} + C_2z_i = u_{2i},$$

whereas  $A_1x_i = u_{1i}$  can be written as

$$B_{11}y_{1i} + B_{12}y_{2i} + C_1z_i = u_{1i}.$$

First, we want to write  $y_2$  in terms of the exogenous variables  $z$ . To that end, denote

$$D = \begin{pmatrix} I & 0 \\ B^{21} & B^{22} \end{pmatrix},$$

where  $B^{21}$  and  $B^{22}$  are the corresponding elements of  $B^{-1}$ . That is,

$$B^{22} = (B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1}; \quad B^{21} = -B^{22}B_{21}B_{11}^{-1}.$$

So, to obtain  $D$  we have only chosen the second set of rows of  $B^{-1}$  and then the first row of  $B^{-1}$  is replaced by  $[I; 0]$ .

Then we obtain after we apply  $D$  that

$$A^* = DA = \begin{pmatrix} B_{11} & B_{12} & C_1 \\ 0 & I_{G-G_1} & -\Pi_2 \end{pmatrix} = \begin{pmatrix} A_1^* \\ A_2^* \end{pmatrix}$$

and

$$\Sigma^* = D\Sigma D' = \begin{pmatrix} \Sigma_{11} & \Sigma_{12}^* \\ \Sigma_{21}^* & \Sigma_{22}^* \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & * \\ B^{21}\Sigma_{11} + B^{22}\Sigma_{21} & \Omega_{22} \end{pmatrix}.$$

Thus, we can now consider the system of equations

$$A^* x_i = u_i^*; \quad A^* = \begin{pmatrix} B_{11} & B_{12} & C_1 \\ 0 & I_{G-G_1} & -\Pi_2 \end{pmatrix}.$$

So, given the above system of equations we can employ the *PMLE*, *MDE*, etc... to estimate the parameters, where  $\Pi_2$ ,  $\Sigma_{21}^*$  and  $\Sigma_{22}^*$  are unrestricted, but  $(B_{11}, B_{12}, C_1, \Sigma_{11})$  are with the constraints we had in the “original” system.

**Definition 16** *The components  $A_1^*, \Sigma_{11}^*$  of  $A^*, \Sigma^*$  minimizing*

$$Q(A^*, \Sigma^*) = \log |\Sigma^*| - 2 \log |B^*| + tr \left( A^* \frac{X'X}{n} A^{*'} \Sigma^{*-1} \right)$$

*are the LIMITED INFORMATION PMLE (LIPMLE) of  $A_1, \Sigma_{11}$ , respectively.*

**Theorem 17** *(Concentrated Likelihood) The LIPMLE of  $A_1, \Sigma_{11}$  is equivalent to*

$$\left( \tilde{A}_1, \tilde{\Sigma}_{11} \right) = \arg \min_{A_1, \Sigma_{11}} \tilde{Q}(A_1, \Sigma_{11}),$$

where

$$\tilde{Q}(A_1, \Sigma_{11}) = \log |\Sigma_{11}| - \log |B_1 \hat{\Omega} B_1'| + \text{tr} \left( A_1 \frac{X'X}{n} A_1' \Sigma_{11}^{-1} \right),$$

and

$$\hat{\Omega} = \frac{1}{n} \left\{ Y' \left( I - Z (Z'Z)^{-1} Z' \right) Y \right\}.$$

**Theorem 18** *If  $\Sigma_{11}$  is unconstrained and only  $A_1$  depends on  $\theta_1$ , then the LIPMLE of  $\theta_1$  is*

$$\tilde{\theta}_1 = \arg \min_{\theta_1 \in \Theta} R(\theta_1),$$

where

$$R(\theta_1) = \frac{\left| A_1 \frac{X'X}{n} A_1' \right|}{\left| B_1 \hat{\Omega} B_1' \right|}.$$

(Compare with Theorem 12).

**Definition 17** The components  $\bar{A}_1, \tilde{A}_1$  of the matrices  $\bar{A}^*, \tilde{A}^*$ , minimizing

$$\begin{aligned} & \frac{1}{n} \text{tr} \left( A^* X' Z (Z' Z)^{-1} Z' X A^{*'} \right) \\ & \frac{1}{n} \text{tr} \left( A^* X' Z (Z' Z)^{-1} Z' X A^{*'} \hat{\Sigma}^{*-1} \right), \end{aligned}$$

where  $\hat{\Sigma}^* = \bar{A}^* \frac{X'X}{n} \bar{A}^{*'}$ , are the subsystem 2SLSE and 3SLSE respectively.

**Theorem 19**

$$\begin{aligned} \bar{A}_1 &= \arg \min_{A_1} \text{tr} \left( A_1 X' Z (Z' Z)^{-1} Z' X A_1' \right); \\ \tilde{A}_1 &= \arg \min_{A_1} \text{tr} \left( A_1 X' Z (Z' Z)^{-1} Z' X A_1' \hat{\Sigma}_{11}^{-1} \right), \end{aligned}$$

where  $\hat{\Sigma}_{11} = \bar{A}_1 \frac{X'X}{n} \bar{A}_1'$ .

**Comments #1** “Single Equation” 3SLSE  $\equiv$  “Single Equation” 2SLSE.

#2 Single Equation 2SLSE differs from system 2SLSE if the latter uses cross-equations restrictions.

#3 Subsystem *3SLSE* is subsystem *3SLSE* on  $A^*x_i = u_i^*$ , which is as efficient as to the *PMLE* based on  $A^*x_i = u_i^*$  ( $\Sigma^*$  unrestricted). The subsystem *3SLSE* is equally efficient to the *LIPMLE*.

## 4 CONSISTENCY OF THE ESTIMATORS

Before we examine the consistency of the different estimators we have proposed, that is the *PMLE*, *MDE*, *2SLSE* or the *3SLSE*, we are going to write again for easy reference a general theorem regarding the consistency of extremum estimators.

**Theorem 20** *Let  $\hat{\alpha}$  be defined as*

$$\hat{\alpha} = \arg \min_{\alpha \in \mathcal{A}} \mathcal{R}(\alpha),$$

where  $\mathcal{A} \subset \mathbb{R}^p$  is a compact set and  $\alpha_0 \in \mathcal{A}$ . Assume that

$$\mathcal{R}(\alpha) - \mathcal{R}(\alpha_0) = \mathcal{S}(\alpha) - \mathcal{T}(\alpha),$$

where  $\mathcal{S}(\alpha)$  is nonstochastic and  $\forall \epsilon > 0, \exists \delta > 0$  such that

$$(a) \quad \inf_{\|\alpha - \alpha_0\| \geq \epsilon} \mathcal{S}(\alpha) \geq \delta > 0 \quad (b) \quad \sup_{\alpha \in \mathcal{A}} |\mathcal{T}(\alpha)| = o_p(1).$$

Then,

$$\hat{\alpha} \xrightarrow{P} \alpha_0.$$

## 4.1 CONSISTENCY OF THE PMLE

Let  $\Psi = (\Pi, \Omega)$  map  $\Theta$  into  $\mathcal{B}$ . We can proceed by finding

$$\widehat{\Psi} = \arg \min_{\mathcal{B}} Q_2(\Psi),$$

then  $\widehat{\theta}$  is the solution to  $\Psi(\widehat{\theta}) = \widehat{\Psi}$ .

Next let's introduce the following regularity conditions.

**A3**  $\Psi = \Psi_0$  implies that  $\Phi = \Phi_0$ .

**A4**  $\Psi = \Psi_0$  implies that  $\theta = \theta_0$ .

**A5** (LLN)

$$\widehat{M} = \frac{Z'Z}{n} \xrightarrow{P} M > 0.$$

Asymptotic uncorrelation between  $Z$  and  $V_0$ ,

$$\widehat{N} = \frac{Z'V_0}{n} \xrightarrow{P} 0$$

and

$$\widehat{O} = \frac{V_0'V_0}{n} \xrightarrow{P} \Omega_0 > 0.$$

**A6**  $\theta_0 \in \Theta$ .

It is implied by A6 that  $\Psi_0 \in \mathcal{B}$  and by A1 that  $\mathcal{B}$  is compact.

**Theorem 21** (*Consistency of the PMLE*)

(a) *If A1, A5, A6 hold, then*

$$\widehat{\Psi} \xrightarrow{P} \Psi_0.$$

(b) *If also A3 holds, then*

$$\widehat{\Phi} \xrightarrow{P} \Phi_0.$$

(c) *If also A2, A4 hold, then*

$$\widehat{\theta} \xrightarrow{P} \theta_0.$$

**Proof.** Denote  $\alpha = \Psi$ ;  $\mathcal{R}(\alpha) = Q_2(\Psi)$ ;  $\mathcal{A} \equiv \mathcal{B}$ . Then,

$$\begin{aligned} \mathcal{R}(\alpha) - \mathcal{R}(\alpha_0) &= Q_2(\Psi) - Q_2(\Psi_0) \\ &= \log |\Omega| - \log |\Omega_0| + \frac{1}{n} \text{tr} \{V'V\Omega^{-1}\} \\ &\quad - \frac{1}{n} \text{tr} \{V_0'V_0\Omega_0^{-1}\}. \end{aligned}$$

Let  $H = \Omega^{-1/2}\Omega_0\Omega^{-1/2}$  and define

$$\mathcal{S}(\alpha) = \text{tr} \{(\Pi - \Pi_0)M(\Pi - \Pi_0)'\Omega^{-1}\} + \text{tr}(H) - \log |H| - G$$

$$\begin{aligned} -\mathcal{T}(\alpha) &= \text{tr} \left\{ (\Pi - \Pi_0) \left( \widehat{M} - M \right) (\Pi - \Pi_0)' \Omega^{-1} \right\} \\ &\quad - 2\text{tr} \left\{ (\Pi - \Pi_0) \widehat{N} \Omega^{-1} \right\} \\ &\quad + \text{tr} \left\{ \left( \widehat{O} - \Omega_0 \right) \left( \Omega^{-1} - \Omega_0^{-1} \right) \right\}, \end{aligned} \tag{4.1}$$

to show (a) and (b) in Theorem 20.

**Proof of (a)** : that is, show

$$\inf_{\|\Psi - \Psi_0\| \geq \epsilon} \mathcal{S}(\alpha) \geq \delta > 0.$$

First  $\|\Psi - \Psi_0\| \geq \epsilon$  means that

$$(I) \quad \|\Pi - \Pi_0\|^2 \geq \epsilon/2 \text{ and/or } (II) \quad \|\Omega - \Omega_0\|^2 \geq \epsilon/2.$$

That is, 1..  $\left\| \begin{pmatrix} \Pi' \\ 0 \end{pmatrix} \right\| = \|\Pi'\|$ ; 2.  $\Psi = \begin{pmatrix} \Pi' \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \Omega \end{pmatrix}$ ; 3. by triangular inequality,  $\|A + B\| \leq \|A\| + \|B\|$ . Thus, if both  $\|A\|$  and  $\|B\|$  are smaller than  $\epsilon/2$  implies that  $\|A + B\| < \epsilon$ .

Next, we have that

$$\text{tr}(H) - \log |H| - G \geq 0$$

because the function  $g(x) = x - \log x - 1$  is continuous for  $x > 0$  and with a

minimum at  $x = 1$ , for which  $g(x = 1) = 0$ , and we have that

$$\text{tr}(H) - \log |H| - G = \sum_{j=1}^G (\lambda_j - \log \lambda_j - 1),$$

where  $\lambda_j$ ,  $j = 1, \dots, G$ , is the  $j$ th eigenvalue of  $H$ . Notice that by definition  $H$  is a positive definite matrix, so that all its eigenvalues are positive. Now

$$\epsilon/2 \leq \|\Omega - \Omega_0\| = \left\| \Omega^{1/2} (I - H) \Omega^{1/2} \right\| \leq \|\Omega\| \|(I - H)\|$$

As  $\|\Omega\|$  is bounded below some positive,

$$\begin{aligned} \{\|H - I\| \geq \epsilon'/2\} &\Leftrightarrow \max_{j=1, \dots, G} |\lambda_j - 1| \geq \epsilon'/2 > 0 \\ &\Leftrightarrow \max_{j=1, \dots, G} (\lambda_j - \log \lambda_j - 1) \geq \epsilon''/2 \\ &\Leftrightarrow \inf_{\mathcal{N}} \sum_{j=1}^G (\lambda_j - \log \lambda_j - 1) \geq \delta > 0. \end{aligned}$$

Next, because we have that

$$\begin{aligned}
\|\Pi - \Pi_0\|^2 &= \left\| \Omega^{1/2} \left( \Omega^{-1/2} (\Pi - \Pi_0) M^{1/2} \right) M^{-1/2} \right\|^2 \\
&\leq \left\| \Omega^{1/2} \right\|^2 \left\| \Omega^{-1/2} (\Pi - \Pi_0) M^{1/2} \right\|^2 \left\| M^{-1/2} \right\|^2 \\
&= \bar{\lambda}(\Omega) \left\| \Omega^{-1/2} (\Pi - \Pi_0) M^{1/2} \right\|^2 \underline{\lambda}^{-1}(M),
\end{aligned}$$

we obtain that

$$\left\| \Omega^{-1/2} (\Pi - \Pi_0) M^{1/2} \right\|^2 \geq \|\Pi - \Pi_0\|^2 \underline{\lambda}(M) \bar{\lambda}^{-1}(\Omega) \geq \delta > 0.$$

Moreover, we have that because for a semi-positive definite matrix, the trace cannot be smaller than the maximum eigenvalue, we can conclude that

$$\begin{aligned}
\mathcal{S}(\alpha) &\geq \left\| \Omega^{-1/2} (\Pi - \Pi_0) M^{1/2} \right\|^2 + \sum_{j=1}^G (\lambda_j - \log \lambda_j - 1) \quad (4.2) \\
&\geq \|\Pi - \Pi_0\|^2 \underline{\lambda}(M) \bar{\lambda}^{-1}(\Omega) + \sum_{j=1}^G (\lambda_j - \log \lambda_j - 1).
\end{aligned}$$

**Proof of (b)** : that is, show that

$$\sup_{\mathcal{B}} |\mathcal{T}(\alpha)| \xrightarrow{P} 0.$$

Regarding the first term on the right of (4.1),

$$\sup_{\mathcal{B}} \left\| (\Pi - \Pi_0) \left( \widehat{M} - M \right) (\Pi - \Pi_0)' \Omega^{-1} \right\| \leq C \left\| \widehat{M} - M \right\|$$

because by compactness of the parameter space and continuity it implies that  $\|\Pi\| < C_1$  and positive definite of  $\Omega$ . But the right side of the last displayed inequality converges to zero in probability by Assumption A5.

The second term on the right of (4.1) is

$$\sup_{\mathcal{B}} \left\| (\Pi - \Pi_0) \widehat{N} \Omega^{-1} \right\| \leq C \left\| \widehat{N} \right\| = o_p(1)$$

by A5. Finally, regarding the third term on the right of (4.1) we have that

$$\left\| \left( \widehat{O} - \Omega_0 \right) \left( \Omega^{-1} - \Omega_0^{-1} \right) \right\| \leq C \left\| \widehat{O} - \Omega_0 \right\| = o_p(1)$$

by the same arguments. This concludes the proof of (2) and hence part (a).

Next we show part (b). Because  $\Psi$  is continuous in  $\Phi$  and by  $A\beta$ ,  $\Psi = \Psi_0$  iff  $\Phi = \Phi_0$ , means that  $\forall \epsilon \geq 0, \exists \delta > 0$  such that

$$\left\| \widehat{\Psi} - \Psi_0 \right\| < \delta \Rightarrow \left\| \widehat{\Phi} - \Phi_0 \right\| < \epsilon.$$

That is, the inverse mapping is continuous. So,

$$\Pr \left\{ \left\| \widehat{\Phi} - \Phi_0 \right\| \geq \epsilon \right\} \leq \Pr \left\{ \left\| \widehat{\Psi} - \Psi_0 \right\| \geq \delta \right\} \rightarrow 0,$$

which proves part (b). Part (c) follows by the same reason. ■

### Remark

1. Assumption  $A2$  very mild. Observe that we do not need differentiability.
2. Assumptions  $A3$  and  $A4$  are identifiability conditions.
3. Assumption  $A5$  is true under several conditions. What we need is not that  $z_i$  and  $v_i$  are uncorrelated but asymptotically uncorrelated. We basically rule out that  $z_i$  contains lagged values and  $v_i$  is correlated.
4.  $\widehat{O} \xrightarrow{P} \Omega_0$  and  $\widehat{M} - M = o_p(1)$  are true even with serial dependence in errors and/or regressors.
5. Assumption  $A6$  is not very restrictive. In fact, in practice, we restrict ourselves to search over  $\mathcal{A}$  on a compact set. The problem is, if any, how to choose it.

## 4.2 CONSISTENCY OF THE MDE

Consider the following conditions.

**B3**  $\Pi = \Pi_0 \Rightarrow A = A_0.$

**B4**  $\Pi = \Pi_0 \Rightarrow \theta = \theta_0.$

**B5**  $\widehat{M} \xrightarrow{p} M > 0; \widehat{N} \xrightarrow{p} 0.$

**B6** The same as A6.

Recall that  $\Pi(\theta)$  is a map  $\Theta \rightarrow \mathcal{B}$  and that

$$Q_2(\widehat{\Pi}, \widehat{\Omega}) = \min_{\mathcal{B}} Q_2(\Pi, \widehat{\Omega}),$$

where

$$\widehat{\Omega} = \frac{1}{n} \left( Y'Y - Y'Z(Z'Z)^{-1}Z'Y \right).$$

Now,  $\widehat{A}$  is such that

$$\widehat{B}\widehat{\Pi} + \widehat{C} = 0; \quad \widehat{\theta} : \widehat{\Pi} = \Pi(\widehat{\theta}).$$

Then, we have the following theorem.

**Theorem 22** (i) *If B1, B2, B5, and B6 hold, then*

$$\widehat{\Pi} \xrightarrow{P} \Pi_0.$$

(ii) *If also B3 holds, then*

$$\widehat{A} \xrightarrow{P} A_0.$$

(iii) *If also B4 holds, then*

$$\widehat{\theta} \xrightarrow{P} \theta_0.$$

### 4.3 CONSISTENCY OF THE 2SLSE AND 3SLSE

Let

$$\frac{1}{n} \text{tr} \left\{ A \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} Z' Z \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix}' A' D \right\}$$

for some positive definite matrix  $D$ .

Let  $A : \Theta \rightarrow \mathcal{B}$  and  $\hat{A}$  denote the value which minimizes the last displayed expression, that is

$$\hat{A} = \arg \min_{\mathcal{B}} \frac{1}{n} \text{tr} \left\{ A \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} Z' Z \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix}' A' D \right\}.$$

Then,

**Theorem 23** *If  $B1, B2, B3, B5, B6$  hold, then*

$$\hat{A} \xrightarrow{P} A_0.$$

For the 2SLSE we have that  $D = I_G$ , whereas for the 3SLSE we have that  $D = \widehat{\Sigma}^{-1}$ . If in addition B4 holds, then

$$\hat{\theta} \xrightarrow{p} \theta_0.$$

**Proof.** We shall examine the general case, that is

$$\mathcal{R}(\alpha) = \frac{1}{n} \text{tr} \left\{ A \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} Z' Z \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix}' A' D \right\},$$

where  $\alpha = A$  and  $D > 0$ .

Recall that when  $D = I_G$ , we have the 2SLSE, whereas for  $D = \widehat{\Sigma}^{-1}$  we obtain the 3SLSE. Now,

$$\begin{aligned} A &: \Theta \rightarrow \mathcal{B}; \\ \widehat{A} &: \mathcal{R}(\widehat{A}) = \min_{\mathcal{B}} \mathcal{R}(A), \\ \widehat{\Pi} &= -\widehat{B}^{-1}\widehat{C}; \quad \widehat{\theta}: A(\widehat{\theta}) = \widehat{A}. \end{aligned}$$

Now,

$$\begin{aligned} A \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} &= A \begin{pmatrix} \Pi_0 + \hat{N}'\hat{M}^{-1} \\ I_K \end{pmatrix} \\ &= B\Pi_0 + C + B\hat{N}'\hat{M}^{-1}. \end{aligned}$$

Also

$$\begin{aligned} A_0 \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} &= B_0\Pi_0 + C_0 + B_0\hat{N}'\hat{M}^{-1} \\ &= B_0\hat{N}'\hat{M}^{-1}. \end{aligned}$$

So, we have that

$$\begin{aligned}
\mathcal{R}(\alpha) - \mathcal{R}(\alpha_0) &= \frac{1}{n} \text{tr} \left\{ A \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} Z' Z \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix}' A' D \right\} \\
&\quad - \frac{1}{n} \text{tr} \left\{ A_0 \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix} Z' Z \begin{pmatrix} \tilde{\Pi} \\ I_K \end{pmatrix}' A_0' D \right\} \\
&= \text{tr} \left\{ \left( B\Pi_0 + C + B\hat{N}'\hat{M}^{-1} \right) \hat{M} \left( B\Pi_0 + C + B\hat{N}'\hat{M}^{-1} \right)' D \right\} \\
&\quad - \text{tr} \left\{ B_0\hat{N}'\hat{M}^{-1}\hat{N}B_0'D \right\}.
\end{aligned}$$

After standard algebra, we obtain then that

$$\mathcal{S}(\alpha) = \text{tr} \left\{ (B\Pi_0 + C) M (B\Pi_0 + C)' D \right\}$$

$$\begin{aligned}
-\mathcal{T}(\alpha) &= \text{tr} \left\{ (B\Pi_0 + C) (\widehat{M} - M) (B\Pi_0 + C)' D \right\} \\
&\quad + 2\text{tr} \left\{ B\widehat{N}' (B\Pi_0 + C)' D \right\} \\
&\quad + \text{tr} \left\{ B\widehat{N}' \widehat{M}^{-1} \widehat{N} B' D \right\} \\
&\quad - \text{tr} \left\{ B_0 \widehat{N}' \widehat{M}^{-1} \widehat{N} B_0' D \right\}.
\end{aligned} \tag{4.3}$$

Thus we need to show that

(1)

$$\inf_{\|\alpha - \alpha_0\| \geq \epsilon} \mathcal{S}(\alpha) \geq \delta > 0,$$

(2)

$$\sup_{\alpha} |\mathcal{T}(\alpha)| \xrightarrow{P} 0.$$

We shall begin showing (1). Now

$$\begin{aligned}
\inf_{\overline{\mathcal{N}}=\{\alpha:\|\alpha-\alpha_0\|>\epsilon\}} \mathcal{S}(\alpha) &\geq \inf_{\overline{\mathcal{N}}=\{\alpha:\|\alpha-\alpha_0\|>\epsilon\}} \left\| D^{1/2} (B\Pi_0 + C) M^{1/2} \right\|^2 \\
&\geq \underline{\lambda}(D) \underline{\lambda}(M) \inf_{\overline{\mathcal{N}}=\{\alpha:\|\alpha-\alpha_0\|>\epsilon\}} \|B\Pi_0 + C\|^2 \\
&\geq F \inf_{\overline{\mathcal{N}}=\{\alpha:\|\alpha-\alpha_0\|>\epsilon\}} \|B\Pi_0 + C\|^2
\end{aligned}$$

because both  $D$  and  $M$  are positive definite,  $\underline{\lambda}(D) \underline{\lambda}(M) \geq F > 0$ . Next, we examine  $\|B\Pi_0 + C\|^2$ . We have that

$$B\Pi_0 + C = (A - A_0) \begin{pmatrix} \Pi_0 \\ I_K \end{pmatrix},$$

and thus

$$\left\| (A - A_0) \begin{pmatrix} \Pi_0 \\ I_K \end{pmatrix} \begin{pmatrix} \Pi_0 \\ I_K \end{pmatrix}' (A - A_0)' \right\| \geq \|(A - A_0)\| \epsilon,$$

for some  $\epsilon > 0$  as the matrix in the middle is p.d.

Next we show (2). The proof is very similar to that of the *PMLE* or *MDE*. Regarding the first term on the right of (4.3),

$$\sup_{\alpha} \left\| (B\Pi_0 + C) \left( \widehat{M} - M \right) (B\Pi_0 + C)' D \right\| \leq F \left\| \widehat{M} - M \right\|$$

by compactness and continuity. But by assumption, the right side of the last displayed inequality converges to zero in probability. Regarding the first term on the right of (4.3), it converges to zero in probability by mimicking the previous argument. Recall that  $\widehat{N} \rightarrow_P 0$  by assumption. So far the contribution due to the third term on the right of (4.3), we have that it is

$$tr \left\{ B\widehat{N}' \left( \widehat{M}^{-1} - M^{-1} \right) \widehat{N}B'D \right\} + tr \left\{ B\widehat{N}' M^{-1} \widehat{N}B'D \right\},$$

so that because  $M > 0$ , we have that  $\widehat{M}^{-1} \rightarrow_P M^{-1}$  by Slutsky's theorem and also because  $\widehat{N} \rightarrow_P 0$ , we conclude that the contribution of the third term on the right of (4.3) is  $o_p(1)$  because

$$\left\| B\widehat{N}' \left( \widehat{M}^{-1} - M^{-1} \right) \widehat{N}B'D \right\| \leq F \left\| \widehat{N} \right\|^2 \left\| \widehat{M}^{-1} - M^{-1} \right\|,$$

as is the fourth term on the right of (4.3). This completes the proof for the general case.

Once we have the proof for a generic  $D$ , we shall give the proof for the  $2SLSE$  and the  $3SLSE$ . For the consistency of the  $2SLSE$  we just choose  $D = I_G$ .

Regarding the  $3SLSE$ , we have that  $D = \widehat{\Sigma}^{-1}$ , where

$$\widehat{\Sigma} = \widehat{A}_{2SLSE} \frac{X'X}{n} \widehat{A}'_{2SLSE}.$$

Now

$$\frac{X'X}{n} = \frac{1}{n} \begin{bmatrix} Y'Y & Y'Z \\ Z'Y & Z'Z \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \Omega_0 + \Pi_0 M \Pi_0' & \Pi_0 M \\ M \Pi_0' & M \end{bmatrix}.$$

Also we have already shown that  $\widehat{A}_{2SLSE} \xrightarrow{P} A_0$ . So,

$$\widehat{\Sigma} \xrightarrow{P} A_0 \begin{bmatrix} \Omega_0 + \Pi_0 M \Pi_0' & \Pi_0 M \\ M \Pi_0' & M \end{bmatrix} A_0' = \Sigma_0$$

because the middle expression is

$$\begin{aligned} & (B_0 \Omega_0 + B_0 \Pi_0 M \Pi_0' + C_0 M \Pi_0'; B_0 \Pi_0 M + C_0 M) A_0' \\ & = B_0 \Omega_0 B_0' = \Sigma_0. \end{aligned}$$

So, we have that the only difference of  $\mathcal{R}(\alpha) - \mathcal{R}(\alpha_0)$  with the one for the general situation is the term

$$\begin{aligned} & tr \left\{ (B\Pi_0 + C) M (B\Pi_0 + C)' \widehat{\Sigma}^{-1} \right\} \\ = & tr \left\{ (B\Pi_0 + C) M (B\Pi_0 + C)' \Sigma_0^{-1} \right\} \\ & + tr \left\{ (B\Pi_0 + C) M (B\Pi_0 + C)' \left( \widehat{\Sigma}^{-1} - \Sigma_0^{-1} \right) \right\}. \end{aligned}$$

The first term on the right of the last displayed equality is identical to that  $\mathcal{S}(\alpha)$  with  $D$  replaced by  $\Sigma_0$ , whereas the second term on the right converges to zero in probability from the previous argument. ■

## 5 ASYMPTOTIC NORMALITY OF THE ESTIMATORS

The next question or issue is as follows. Knowing that  $\hat{\theta}$  is a consistent estimator, we now wish to give conditions under which  $\hat{\theta}$  is asymptotically normal.

Recall that we shall abbreviate  $Q_n(\theta)$  by  $Q(\theta)$ . The following is a general result on the asymptotic normality of the extremum estimators.

**Theorem 24** *Assume*

- (i)  $\theta_0$  is an interior point of the compact set  $\Theta$ .
- (ii)  $Q(\theta)$  is a twice continuously differentiable function in a neighbourhood of  $\theta_0$
- (iii)

$$n^{1/2} \frac{\partial Q(\theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N}(0, D) \tag{2}$$

$$\frac{\partial^2 Q(\tilde{\theta})}{\partial \theta \partial \theta'} \xrightarrow{P} E > 0 \tag{3}$$

for all  $\tilde{\theta}$  such that  $\tilde{\theta} \xrightarrow{P} \theta_0$  and where  $E$  is a positive definite matrix.

(iv)  $\hat{\theta} \xrightarrow{P} \theta_0$ .

Then,

$$n^{1/2} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, E^{-1} D E^{-1} \right).$$

**Remark 5** The condition (i) guarantees that  $\frac{\partial Q(\hat{\theta})}{\partial \theta} = 0$ .

**Remark 6** In many situations such as when we have that  $\hat{\theta}$  is the Maximum-Likelihood or the Nonlinear least squares estimators with independent errors, we have that  $E = D$ . In this case, we would obtain that

$$n^{1/2} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, E^{-1} \right).$$

However, in many other situations, for instance the Instrumental variables estimators, we have that  $D \neq E$ .

## 5.1 PMLE

To apply the above theorem to our estimators we need to strengthen our regularity conditions.

**A7**  $\theta_0$  is an interior point of  $\Theta$ .

**A8** We have that

$$n^{1/2} \begin{pmatrix} \text{vec}(\widehat{N}) \\ \text{vec}(\widehat{O} - \Omega_0) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, L).$$

**A9**  $\Phi$  is a twice continuously differentiable function in  $\theta$ .

**A10**  $\theta_0$  is a regular point of  $\begin{bmatrix} P \\ W \end{bmatrix}$ , where

$$P = \frac{\partial}{\partial \theta'} \text{vec}(\Pi'); \quad W = \frac{\partial}{\partial \theta'} \text{vec}(\Omega).$$

**Notation:** Let

$$\Pi_i = \frac{\partial \Pi}{\partial \theta_i}; \quad \Omega_i = \frac{\partial \Omega}{\partial \theta_i}$$

$$h_i = \text{vec} \left[ \left( \Pi'_i; \frac{1}{2} \Omega^{-1} \Omega_i \right) \Omega^{-1} \right]_{\theta_0},$$

$$H' = (h_1, \dots, h_p) = \begin{bmatrix} \Omega^{-1} \otimes I_K & 0 \\ 0 & \frac{1}{2} (\Omega^{-1} \otimes \Omega^{-1}) \end{bmatrix} \begin{bmatrix} P \\ W \end{bmatrix}_{\theta_0}.$$

Moreover, denoting

$$e_{ij} = \text{tr} \left[ (2\Pi_i M \Pi'_j + \Omega_i \Omega^{-1} \Omega_j) \Omega^{-1} \right]_{\theta_0},$$

we have that

$$\begin{aligned} E &= [e_{ij}]_{i,j=1,\dots,p} \\ &= [P'; W'] \begin{bmatrix} \Omega^{-1} \otimes M & 0 \\ 0 & \frac{1}{2} (\Omega^{-1} \otimes \Omega^{-1}) \end{bmatrix} \begin{bmatrix} P \\ W \end{bmatrix}_{\theta_0}. \end{aligned}$$

1. The matrix  $H$  comes from the first derivative of the objective function, i.e.  $\partial Q(\theta) / \partial \theta$ ,
2.  $E$  is the second derivative of the objective function, i.e.

$$\left\{ \partial^2 Q(\theta) / \partial \theta_i \partial \theta_j \right\}_{i,j=1,\dots,p}$$

after using the fact that

$$\text{tr}(ABCD) = \text{vec}'(C)(D \otimes B')\text{vec}(A').$$

Also, we assume *A1* and *A7*, that is the compactness and asymptotic identifiability respectively.

**Theorem 25** Let  $\widehat{\theta}$  be the PMLE of  $\theta$ . Assume that A1 – A10 hold. Then,

$$n^{1/2} \left( \widehat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, E_0^{-1} D_0 E_0^{-1} \right); \quad D = HLH'.$$

Some conditions are redundant. For instance, A6 is implied by A7.

**Proof.** First, we shall to show that

$$(a) \widehat{\theta} \xrightarrow{P} \theta_0 \text{ and } (b) E > 0. \tag{5.2}$$

As (a) is already shown, we shall show part (b) of (5.2). We first observe that A4 & A10 imply that

$$\text{rank} \begin{bmatrix} P \\ W \end{bmatrix}_{\theta_0} = p,$$

which in turn implies that  $E > 0$ .

So, it remains to show that conditions in (iii) of Theorem 24 are satisfied. We shall begin with (2) For that purpose, we shall show that <sup>3</sup>

$$\frac{1}{n^{1/2}} \frac{\partial}{\partial \theta} Q(\theta_0) \xrightarrow{d} \mathcal{N} \left( 0, D_0 \right); \quad D_0 = HLH' |_{\theta_0}.$$

---

<sup>3</sup>Review of matrix calculus: for  $A = A(\theta)$

Denoting

$$\begin{aligned} \frac{1}{n^{1/2}} Q_{i0} &= \frac{1}{n^{1/2}} \frac{\partial Q(\theta_0)}{\partial \theta_i} \\ &= -\text{vec}' \left[ \left( \Pi'_i; \frac{1}{2} \Omega^{-1} \Omega_i \right) \Omega^{-1} \right] S |_{\theta_0}, \end{aligned}$$

where

$$S = n^{1/2} \begin{bmatrix} \text{vec}(\widehat{N}) \\ \text{vec}(\widehat{O} - \Omega_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, L).$$

- 
1.  $\frac{\partial}{\partial \theta} \ln |A| = \left\{ \frac{\partial}{\partial \theta} \text{vec}'(A) \right\} \text{vec}(A^{-1})$ .
  2.  $\frac{\partial}{\partial \theta} \text{vec}'(A^{-1}) = - \left\{ \frac{\partial}{\partial \theta} \text{vec}'(A) \right\} (A^{-1} \otimes A'^{-1})$ .
  3.  $\frac{\partial}{\partial \theta} \text{tr}(AB) = \left\{ \frac{\partial}{\partial \theta} \text{vec}'(A) \right\} \text{vec}(B')$ .
  4.  $\frac{\partial}{\partial \theta} \text{tr}(A' B A C) = 2 \frac{\partial \text{vec}'(A)}{\partial \theta} (C \otimes B') \text{vec}(A)$ .
  5.  $\frac{\partial^2}{\partial \theta \partial \theta'} \text{tr}(A' B A C) = 2 \frac{\partial \text{vec}'(A)}{\partial \theta} (C \otimes B') \frac{\partial \text{vec}(A)}{\partial \theta} + M$ , where  $M$  has  $(i, j)$ -th element  $2 \text{tr} \left( A' B \frac{\partial^2 A}{\partial \theta_i \partial \theta_j} C \right)$ .

So, we can conclude that

$$\frac{1}{n^{1/2}} \frac{\partial Q(\theta_0)}{\partial \theta} = -HS \xrightarrow{d} \mathcal{N}(0, HLH').$$

This completes the proof of (2). Next we show (3). But this is the case because

$$\frac{1}{n} \frac{\partial^2 Q(\tilde{\theta})}{\partial \theta_i \partial \theta_j} \xrightarrow{P} e_{ij}$$

because  $\tilde{\theta}$  is an intermediate point between  $\hat{\theta}$  and  $\theta_0$  which implies that  $\tilde{\theta} \rightarrow_P \theta_0$ .

■

**Remark 7**  $E > 0$  is sufficient for identification and necessary if A10 holds.

**Remark 8** Condition A8 holds true in many situations. For both stochastic and nonstochastic  $z_t$ , lagged  $y_t$  and serially correlated  $v_t$ , but not both at the same time.

### 5.1.1 WHEN $\Pi$ AND $\Omega$ ARE FUNCTIONALLY UNRELATED

We shall assume that

$$\theta = \begin{pmatrix} \theta_{\Pi} \\ \theta_{\Omega} \end{pmatrix}; \quad \Pi = \Pi(\theta_{\Pi}); \quad \Omega = \Omega(\theta_{\Omega}).$$

In this case the matrix  $[P', W']'$  is block diagonal, that is

$$\begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} = \frac{\partial}{\partial \theta'} \text{vec}(\Pi', \Omega)$$

and also we have that

$$E = \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix},$$

where

$$\begin{aligned} E_1 &= P_1' (\Omega^{-1} \otimes M) P_1 \\ E_2 &= \frac{1}{2} P_2' (\Omega^{-1} \otimes \Omega^{-1}) P_2. \end{aligned}$$

Thus, we have that

$$\begin{aligned} H' &= \begin{pmatrix} (\Omega^{-1} \otimes I_K) & 0 \\ 0 & \frac{1}{2} (\Omega^{-1} \otimes \Omega^{-1}) \end{pmatrix} \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \\ &= \begin{pmatrix} (\Omega^{-1} \otimes I_K) P_1 & 0 \\ 0 & \frac{1}{2} (\Omega^{-1} \otimes \Omega^{-1}) P_2 \end{pmatrix} \end{aligned}$$

and then that

$$n^{1/2} \begin{pmatrix} \widehat{\theta}_\Pi - \theta_{0,\Pi} \\ \widehat{\theta}_\Omega - \theta_{0,\Omega} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0; E^{-1} D E^{-1}),$$

where in this case we have that

$$E^{-1} D E^{-1} = \begin{bmatrix} E_1^{-1} H_1 L_{11} H_1' E_1^{-1} & E_1^{-1} H_1 L_{12} H_2' E_2^{-1} \\ & E_2^{-1} H_2 L_{22} H_2' E_2^{-1} \end{bmatrix}.$$

But,

$$L_{11} = \Omega_0 \otimes M$$

is the asymptotic covariance matrix of  $n^{1/2}vec(\widehat{N})$ . Then,

$$\begin{aligned} & n^{1/2}(\widehat{\theta}_{\Pi} - \theta_{0,\Pi}) \xrightarrow{d} \mathcal{N}\left(0, (P_1'(\Omega^{-1} \otimes M)P_1)^{-1} |_{\theta_{0,\Pi}}\right) \\ \equiv & \mathcal{N}\left(0, \left(\widetilde{P}_1' \left(\Sigma^{-1} \otimes \begin{pmatrix} \Pi \\ I \end{pmatrix} M \begin{pmatrix} \Pi \\ I \end{pmatrix}'\right) \widetilde{P}_1\right)^{-1} |_{\theta_{0,\Pi}}\right), \end{aligned}$$

where

$$\widetilde{P}_1 = \frac{\partial}{\partial \theta_{\Pi}'} vec(A').$$

Why? Because  $B\Pi + C = 0$ , we have that  $\begin{pmatrix} \Pi \\ I \end{pmatrix}' A' = 0$  so that

$$\frac{\partial}{\partial \theta_i} \begin{pmatrix} \Pi \\ I \end{pmatrix}' A' = \begin{pmatrix} \Pi \\ I \end{pmatrix}' \frac{\partial}{\partial \theta_i} A' + \frac{\partial \Pi'}{\partial \theta_i} B' = 0.$$

Then,

$$\frac{\partial \Pi'}{\partial \theta_i} B' = - \begin{pmatrix} \Pi \\ I \end{pmatrix}' \frac{\partial}{\partial \theta_i} A'$$

and

$$\begin{aligned}
 -\text{vec}\left(\frac{\partial \Pi'}{\partial \theta'_i} B'\right) &= \text{vec}\left(\left(\begin{pmatrix} \Pi \\ I \end{pmatrix}' \frac{\partial}{\partial \theta'_i} A'\right)\right) \\
 &= \left(I \otimes \left(\begin{pmatrix} \Pi \\ I \end{pmatrix}'\right)\right) \text{vec}\left(\frac{\partial}{\partial \theta'_i} A'\right).
 \end{aligned}$$

So, we conclude that

$$-(B \otimes I) P_1 = \left(I \otimes \left(\begin{pmatrix} \Pi \\ I \end{pmatrix}'\right)\right) \tilde{P}_1.$$

From here

$$P_1 = \left(B^{-1} \otimes \left(\begin{pmatrix} \Pi \\ I \end{pmatrix}'\right)\right) \tilde{P}_1$$

and thus

$$\begin{aligned}
& P_1' (\Omega^{-1} \otimes M) P_1 \\
= & \tilde{P}_1' \left( B'^{-1} \otimes \begin{pmatrix} \Pi \\ I \end{pmatrix} \right) (\Omega^{-1} \otimes M) \left( B^{-1} \otimes \begin{pmatrix} \Pi \\ I \end{pmatrix} \right)' \tilde{P}_1 \\
= & \tilde{P}_1' \left( \Sigma^{-1} \otimes \begin{pmatrix} \Pi \\ I \end{pmatrix} M \begin{pmatrix} \Pi \\ I \end{pmatrix}' \right) \tilde{P}_1.
\end{aligned}$$

If  $u_t$  is Gaussian we have that  $L_{12} = 0$  so that  $\hat{\theta}_\Pi$  and  $\hat{\theta}_\Omega$  are asymptotically uncorrelated. Before we examine the asymptotic normality of the *3SLS* or *MD* estimators, we are going to see a general theorem about the order of magnitude of the difference of estimators.

## 5.2 ASYMPTOTIC EQUIVALENCE OF ESTIMATES

We begin with a proposition. Let  $F_n(\theta)$  and  $H_n(\theta)$  be two loss functions for  $\theta$ .

- Let  $\widehat{\theta}_n$  be such that  $f_n(\widehat{\theta}_n) = 0$ , where  $f_n(\theta)$  is the first order conditions of  $F_n(\theta)$ , that is  $f_n(\theta) = \partial F_n(\theta) / \partial \theta$ .
- Let  $\widetilde{\theta}_n$  be such that  $g_n(\widetilde{\theta}_n) = 0$ , where  $g_n(\theta)$  is the first order conditions of  $H_n(\theta)$ , that is  $g_n(\theta) = \partial H_n(\theta) / \partial \theta$ .

Then, we have the following result

**Theorem 26** *Let  $\mathcal{N}_\epsilon = \{\theta : \|\theta - \theta_0\| < \epsilon\}$  for some  $\theta_0$  and assume that*

(i)  $\widehat{\theta}_n \rightarrow_P \theta_0$ .

(ii)  $\widetilde{\theta}_n \rightarrow_P \theta_0$ .

(iii) *For some  $\epsilon > 0$ ,  $g_n(\theta)$  has a first continuous derivative  $G_n(\theta) = \partial g_n(\theta) / \partial \theta$  for  $\theta \in \mathcal{N}_\epsilon$ , and such that  $\forall \delta > 0$*

$$\lim_{\epsilon \rightarrow 0} \lim_{n \rightarrow \infty} \Pr \left\{ \sup_{\theta \in \mathcal{N}_\epsilon} \|G_n(\theta) - G_n(\theta_0)\| > \delta \right\} = 0$$

$$G_n(\theta_0) = G + o_p(1)$$

with  $|G| \neq 0$ .

Then, we have that<sup>4</sup>

$$\begin{aligned}\widehat{\theta}_n - \widetilde{\theta}_n &= O_p\left(\left\|f_n\left(\widehat{\theta}_n\right) - g_n\left(\widehat{\theta}_n\right)\right\|\right) \\ &= O_p\left(\left\|g_n\left(\widehat{\theta}_n\right)\right\|\right).\end{aligned}$$

**Proof.** Choose  $\epsilon > 0$  that satisfies (iii). Because both  $\widehat{\theta}_n$  and  $\widetilde{\theta}_n$  are consistent, then we have that with probability close to one for some  $n_0$  we have that  $\forall n \geq n_0$

$$\widehat{\theta}_n := \widehat{\theta} \in \mathcal{N}_\epsilon \text{ and } \widetilde{\theta}_n := \widetilde{\theta} \in \mathcal{N}_\epsilon.$$

---

<sup>4</sup>Obviously, we could have written the statement of the theorem as

$$\widehat{\theta}_n - \widetilde{\theta}_n = O_p\left(\left\|f_n\left(\widetilde{\theta}_n\right)\right\|\right).$$

But from **(iii)** we also have that there exists  $\bar{G}_n$  such that

$$\begin{aligned} g_n(\hat{\theta}) - f_n(\hat{\theta}) &= g_n(\hat{\theta}) \\ &= g_n(\tilde{\theta}) + \bar{G}_n(\hat{\theta} - \tilde{\theta}) \\ &= \bar{G}_n(\hat{\theta} - \tilde{\theta}), \end{aligned}$$

where  $\|\bar{G}_n - G_n(\theta_0)\| = O_p(\sup_{\theta \in \mathcal{N}_\epsilon} \|G_n(\theta) - G_n(\theta_0)\|) = o_p(1)$  as  $n \rightarrow \infty$  and  $\epsilon \rightarrow 0$ . Hence,

$$\begin{aligned} \bar{G}_n &= (\bar{G}_n - G_n(\theta_0)) + (G_n(\theta_0) - G) + G \\ &= G + o_p(1). \end{aligned}$$

Therefore, we have that

$$\begin{aligned} \hat{\theta} - \tilde{\theta} &= \bar{G}_n^{-1} g_n(\hat{\theta}) \\ &= (G + o_p(1))^{-1} g_n(\hat{\theta}) \end{aligned}$$

and thus

$$\hat{\theta} - \tilde{\theta} = O_p\left(\|g_n(\hat{\theta})\|\right).$$

This completes the proof. ■

Basically the previous proposition shows that to examine what it is the order of magnitude between the difference of two rival estimators of a parameter, say  $\theta_0$ , it suffices to obtain the order of magnitude of the *FOC* but when evaluated at the other estimator.

In what follows, we shall assume that  $\Sigma$  is unrestricted and that  $\theta$  parameterizes the matrix  $A$  alone. In addition, let's introduce the following conditions, some of which override previous ones.

**B7**  $\theta_0$  is an interior point of  $\Theta$ .

**B8**  $n^{1/2}vec(\widehat{N}) \xrightarrow{d} \mathcal{N}(0, L_1)$ .

**B9**  $\widehat{O} \xrightarrow{p} \Omega_0 > 0$ .

**B10**  $A$  is a twice continuously differentiable function in  $\theta$ .

**B11**  $\theta_0$  is a regular point of

$$P = \frac{\partial}{\partial \theta'} vec(\Pi').$$

**Theorem 27** *Assuming B1 – B11, we have that*

$$\begin{aligned} (a) \quad \widehat{\theta}_{PMLE} - \widehat{\theta}_{MDE} &= O_p\left(n^{-\frac{3}{2}}\right), \\ (b) \quad \widehat{\theta}_{PMLE} - \widehat{\theta}_{3SLSE} &= O_p\left(n^{-1}\right) \\ (c) \quad \widehat{\theta}_{MDE} - \widehat{\theta}_{3SLSE} &= O_p\left(n^{-1}\right). \end{aligned}$$

**Proof.** We shall prove (a) and (c) only. We leave (b) as an exercise. We have shown that  $\widehat{\theta}_{PMLE}$ ,  $\widehat{\theta}_{MDE}$  and  $\widehat{\theta}_{3SLSE}$  are all consistent estimators for  $\theta_0$ .

Also, by a slight modification of Theorem 25, we obtain that

$$\widehat{\theta}_{PMLE} - \theta_0 = O_p\left(n^{-1/2}\right).$$

To show part (a), we shall make use of Theorem 26. To that end, denote  $\widehat{\theta}_n = \widehat{\theta}_{PMLE}$  and  $\widetilde{\theta}_n = \widehat{\theta}_{MDE}$ . Also denote

$$\begin{aligned} f_{n,i}(\theta) &= \frac{\partial}{\partial \theta_i} Q_n(\theta) \\ g_{n,i}(\theta) &= \frac{\partial}{\partial \theta_i} Q_{n,2}(\Pi; \widehat{\Omega}). \end{aligned}$$

Then, we have by definition that  $f_n(\hat{\theta}_n) = 0$  and  $g_n(\tilde{\theta}_n) = 0$ , where<sup>5</sup>

$$\begin{aligned} f_{n,i}(\theta) &= \text{tr}(A_i R' \Sigma^{-1}(A)) \\ g_{n,i}(\theta) &= \text{tr}\left(A_i R' B^{-1} \hat{\Omega}^{-1} B'^{-1}\right) \end{aligned}$$

---

<sup>5</sup>This footnote is to indicate how we can obtain the function  $f_n(\theta)$ . We first remember that the objective function is by Theorem 12,

$$\frac{1}{2} \log \left| \frac{V'V}{n} \right|.$$

So,

$$\frac{\partial}{\partial \theta_i} \frac{1}{2} \log \left| \frac{V'V}{n} \right| = \text{tr} \left( \frac{V'_i V}{n} \hat{O}^{-1} \right).$$

Next,  $V_i = \partial V / \partial \theta_i = -Z \Pi'_i$ , so it implies that

$$\begin{aligned} \frac{\partial}{\partial \theta_i} \frac{1}{2} \log \left| \frac{V'V}{n} \right| &= -\text{tr} \left( \Pi_i \frac{Z'V}{n} \left( \frac{V'V}{n} \right)^{-1} \right) \\ &= -\text{tr} \left( \Pi_i \frac{Z'V}{n} B' \Sigma^{-1}(A) B \right) \end{aligned} \tag{5.3}$$

since  $V' = B^{-1} A X'$  or

$$\begin{aligned} V &= Y - Z \Pi' = Y + Z C' B'^{-1} \\ &= X A' B'^{-1}. \end{aligned}$$

But,  $B \Pi + C = 0$ , which implies that

$$\frac{\partial}{\partial \theta_i} (B \Pi + C) = B_i \Pi + C_i + B \Pi_i$$

with

$$\begin{aligned} A_i &= \frac{\partial}{\partial \theta_i} A; & R &= A \frac{X'Z}{n} (\Pi', I_K) \\ \Sigma(A) &= A \frac{X'X}{n} A'. \end{aligned}$$

Therefore,

$$f_{n,i}(\widehat{\theta}_n) - g_{n,i}(\widehat{\theta}_n) = \text{tr} \left( L_{PMLE,i} \left( \widehat{\Omega}_{PMLE}^{-1} - \widehat{\Omega}^{-1} \right) \right),$$

---

or that

$$-\Pi_i = B^{-1} B_i \Pi + B^{-1} C_i.$$

So, the right side of (5.3) becomes

$$\begin{aligned} & \text{tr} \left( B^{-1} (B_i \Pi + C_i) \frac{Z'V}{n} B' \Sigma^{-1}(A) B \right) \\ &= \text{tr} \left( (B_i \Pi + C_i) \frac{Z'V}{n} B' \Sigma^{-1}(A) \right). \end{aligned}$$

where

$$\begin{aligned}
L_{PMLE,i} &= B'^{-1}A_iR'B^{-1} \Big|_{\theta=\hat{\theta}_n} \\
\hat{\Omega}_{PMLE} &= \frac{1}{n} \left( Y - Z\hat{\Pi}'_{PMLE} \right)' \left( Y - Z\hat{\Pi}'_{PMLE} \right) \\
\hat{\Pi}_{PMLE} &= \Pi \left( \hat{\theta}_n \right).
\end{aligned}$$

Then, we obtain that

$$L_{PMLE,i} = O_p \left( n^{-1/2} \right),$$

because

$$\begin{aligned}
\hat{R} &= \hat{A} \frac{X'Z}{n} \left( \hat{\Pi}', I \right) \\
&= \left( \hat{A} - A_0 \right) \frac{X'Z}{n} \left( \hat{\Pi}', I \right) + A_0 \frac{X'Z}{n} \left( \hat{\Pi}', I \right) \\
&= O_p \left( n^{-1/2} \right) + \frac{U'Z}{n} \left( \hat{\Pi}', I \right) \\
&= O_p \left( n^{-1/2} \right).
\end{aligned}$$

In addition, because

$$\widehat{\Pi}_{PMLE} - \Pi_0 = O_p\left(n^{-1/2}\right); \quad \widetilde{\Pi} - \Pi_0 = O_p\left(n^{-1/2}\right),$$

it implies that

$$\begin{aligned} \widehat{\Omega}_{PMLE} - \widehat{\Omega} &= \left(\widehat{\Pi}_{PMLE} - \widetilde{\Pi}\right) \frac{Z'Z}{n} \left(\widehat{\Pi}_{PMLE} - \widetilde{\Pi}\right)' \\ &= O_p\left(n^{-1}\right) \end{aligned}$$

where  $\widetilde{\Pi}$  is the *LSE* of  $\Pi$ .

Thus,

$$\widehat{\Omega}_{PMLE}^{-1} - \widehat{\Omega}^{-1} = \widehat{\Omega}^{-1} \left(\widehat{\Omega} - \widehat{\Omega}_{PMLE}\right) \widehat{\Omega}_{PMLE}^{-1} = O_p\left(n^{-1}\right)$$

since  $\Omega_0 > 0$  and both  $\widehat{\Omega}$  and  $\widehat{\Omega}_{PMLE}$  are consistent estimators of  $\Omega_0$  which implies that  $\widehat{\Omega}^{-1}$  and  $\widehat{\Omega}_{PMLE}^{-1}$  are  $O_p(1)$ . Then, we have concluded the proof of part (a) of the theorem because

$$f_{n,i}\left(\widehat{\theta}_n\right) - g_{n,i}\left(\widehat{\theta}_n\right) = O_p\left(n^{-1/2}n^{-1}\right) = O_p\left(n^{-3/2}\right).$$

The proof of part (c) is trivial, when we take part (b) as holding true. Indeed, by definition

$$\begin{aligned}\widehat{\theta}_{MDE} - \widehat{\theta}_{3SLSE} &= \left(\widehat{\theta}_{MDE} - \widehat{\theta}_{PMLE}\right) + \left(\widehat{\theta}_{PMLE} - \widehat{\theta}_{3SLSE}\right) \\ &= O_p\left(n^{-3/2}\right) + O_p\left(n^{-1}\right)\end{aligned}$$

by parts (a) and (b) respectively. ■

The previous theorem shows that the *MDE* is the closest estimator to the “IDEAL” and closer than the *3SLSE*.

**Remark 9** *We do not say anything about the 2SLSE because the latter estimator does not have the same asymptotic distribution as the MDE, PMLE or 3SLSE. So, it does not make any sense to study how close the latter estimators are with respect to the 2SLSE.*

*Moreover, the last theorem can be employed to obtain the CLT for the MDE and/or 3SLSE quite trivially.*

**Theorem 28** *Under B1 – B11, we have that*

$$n^{1/2} \left( \widehat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, E_0^{-1} D_0 E_0^{-1} \right),$$

where

$$\begin{aligned} D_0 &= P_0' \left( \Omega_0^{-1} \otimes I_K \right) L_1 \left( \Omega_0^{-1} \otimes I_K \right) P_0 \\ E_0 &= P_0' \left( \Omega_0^{-1} \otimes M \right) P_0 \end{aligned}$$

for  $\widehat{\theta} = \widehat{\theta}_{PMLE}$ , or  $\widehat{\theta}_{MDE}$  or  $\widehat{\theta}_{3SLSE}$ .

Comparing the asymptotic variance with the one in Theorem 25, one notices that it is as if the true  $\Omega$  were known.

**Proof.** A slight modification of Theorem 25 would yield that

$$n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, E_0^{-1} D_0 E_0^{-1} \right).$$

On the other hand, by Theorem 27,

$$\begin{aligned}
 n^{1/2} \left( \widehat{\theta}_{MDE} - \theta_0 \right) &= n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) + n^{1/2} \left( \widehat{\theta}_{MDE} - \widehat{\theta}_{PMLE} \right) \\
 &= n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) + O_p \left( n^{-1} \right) \\
 &= n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) + o_p \left( 1 \right).
 \end{aligned}$$

So, we conclude by standard arguments that

$$n^{1/2} \left( \widehat{\theta}_{MDE} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, E_0^{-1} D_0 E_0^{-1} \right).$$

Now, by Theorem 27 again,

$$\begin{aligned}
 n^{1/2} \left( \widehat{\theta}_{3SLSE} - \theta_0 \right) &= n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) + n^{1/2} \left( \widehat{\theta}_{3SLSE} - \widehat{\theta}_{PMLE} \right) \\
 &= n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) + O_p \left( n^{-1/2} \right) \\
 &= n^{1/2} \left( \widehat{\theta}_{PMLE} - \theta_0 \right) + o_p \left( 1 \right).
 \end{aligned}$$

So, by standard arguments we have the desired result. This completes the proof of the theorem. ■

Similarly, we can deduce the asymptotic distribution of  $\widehat{\Phi}$  and  $\widehat{\Psi}$  which correspond to  $\widehat{\theta}_{PMLE}$ ,  $\widehat{\theta}_{MDE}$  and  $\widehat{\theta}_{3SLSE}$ . We are going to discuss this in a general framework.

### 5.3 DELTA METHOD

Consider a  $q \times 1$  vector of parameters  $\phi(\theta)$ . and also denote its first derivative by  $F$ , that is  $F(\theta)$ .

**Lemma 29** *Suppose that  $F(\theta) = \frac{\partial \phi(\theta)}{\partial \theta'}$  is continuous at  $\theta_0$  and*

$$n^{1/2} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N}(0, J).$$

*Then,*

$$n^{1/2} \left( \phi \left( \hat{\theta} \right) - \phi \left( \theta_0 \right) \right) \xrightarrow{d} \mathcal{N} \left( 0, F J F' \right), \quad (5.3)$$

*where  $F = F(\theta_0)$ .*

**Proof.** The mean value theorem implies that the left side of (5.3) is

$$F \left( \tilde{\theta} \right) n^{1/2} \left( \hat{\theta} - \theta_0 \right),$$

where  $\tilde{\theta}$  is an intermediate point between  $\hat{\theta}$  and  $\theta_0$ . But we have assumed that  $F(\theta)$  is continuous at  $\theta_0$ . So, by Slutsky's theorem, we have that

$$F \left( \tilde{\theta} \right) \xrightarrow{P} F \left( \theta_0 \right),$$

because  $\tilde{\theta} \rightarrow_P \theta_0$ . Therefore, by standard arguments, we conclude that

$$\begin{aligned} n^{1/2} \left( \phi(\hat{\theta}) - \phi(\theta_0) \right) &= F(\theta_0) n^{1/2} (\hat{\theta} - \theta_0) \\ &\quad + \left( F(\tilde{\theta}) - F(\theta_0) \right) n^{1/2} (\hat{\theta} - \theta_0) \\ &= F(\theta_0) n^{1/2} (\hat{\theta} - \theta_0) + o_p(1). \end{aligned}$$

From here we obtain (5.3) by a simple application of Cramer's theorem. ■

**Remark 10** *The identifiability of  $\phi$  means that  $F$  is of full column rank and the number of rows is bigger than equal to that of columns. In terms of the structural form parameters, the distribution is **singular**, that is  $FJF'$  has a determinant equal to zero as is the case with the reduced form parameters if the system of equations is **overidentifiable**, whereas it is **nonsingular** if it is **just-identifiable**. We can expect this because*

$$\phi(\theta) = \text{vec}(\Phi(\theta)) \quad \text{or} \quad \phi(\theta) = \text{vec}(\Psi(\theta)).$$

*For instance, the three structural form parameters depend on two underlying pa-*

rameters. That is,

$$\begin{aligned}\phi_1 &= \theta_1 + \theta_2 \\ \phi_2 &= \theta_1 - \theta_2 \\ \phi_3 &= 2\theta_1 + 3\theta_2.\end{aligned}$$

Obviously,  $\phi_3$  is a linear combination of  $\phi_1$  and  $\phi_2$  which implies that the asymptotic Variance-Covariance matrix of  $\widehat{\phi}_1, \widehat{\phi}_2$  and  $\widehat{\phi}_3$  is SINGULAR.

## 5.4 LAGRANGE MULTIPLIER METHODS

If  $\phi(\theta)$  is only implicitly defined by constraints on  $\phi$ , we can obtain an equivalent result without having to ‘leave out’ the constraints.

Let

$$\widehat{\phi} = \arg \min_{\phi \in \mathcal{A}} Q(\phi),$$

where  $\mathcal{A} \subset \Theta$  is such that  $w(\phi) = 0$  for all  $\phi \in \mathcal{A}$ . Now, the standard way to obtain  $\widehat{\phi}$  is by computing the lagrangean

$$\mathcal{L}(\phi, \lambda) = Q(\phi) + \lambda'w(\phi)$$

and then that

$$\left(\widehat{\phi}', \widehat{\lambda}\right) = \arg \min_{\phi, \lambda} \mathcal{L}(\phi, \lambda).$$

Assuming the consistency of the estimators, we look at their asymptotic distribution.

Now, the F.O.C. are

$$\begin{aligned} 0 &= n^{1/2} \frac{\partial}{\partial \phi} \mathcal{L}(\phi, \lambda) \Big|_{\widehat{\phi}, \widehat{\lambda}} = n^{1/2} \frac{\partial}{\partial \phi} Q(\widehat{\phi}) + n^{1/2} \widehat{W} \widehat{\lambda} \\ &= n^{1/2} \frac{\partial}{\partial \phi} Q(\phi_0) + \widetilde{K} n^{1/2} (\widehat{\phi} - \phi_0) + n^{1/2} \widehat{W} \widehat{\lambda}, \end{aligned}$$

by the mean value theorem, and where

$$\begin{aligned} \widehat{W} &= W(\widehat{\phi}); \quad W(\phi) = \frac{\partial}{\partial \phi} w'(\phi) \\ \widetilde{K} &= \frac{\partial Q(\widetilde{\phi})}{\partial \phi \partial \phi'} = \frac{\partial Q(\phi_0)}{\partial \phi \partial \phi'} + o_p(1) \\ &= K_0 + o_p(1). \end{aligned}$$

Also

$$\begin{aligned} 0 &= n^{1/2} \frac{\partial}{\partial \lambda} \mathcal{L}(\phi, \lambda) \Big|_{\widehat{\phi}, \widehat{\lambda}} = n^{1/2} w(\widehat{\phi}) \\ &= n^{1/2} w(\phi_0) + n^{1/2} \widetilde{W}' (\widehat{\phi} - \phi_0) \\ &= W'_0 n^{1/2} (\widehat{\phi} - \phi_0) + o_p(1), \end{aligned}$$

because  $W(\phi)$  is a continuous function around  $\phi_0$ , so that

$$\widehat{\phi} \xrightarrow{P} \phi_0 \Rightarrow \widetilde{W}' \xrightarrow{P} W_0.$$

Thus, we have the following system

$$\begin{aligned} o_p(1) &= n^{1/2} \frac{\partial}{\partial \phi} Q(\phi_0) + K_0 n^{1/2} (\widehat{\phi} - \phi_0) + n^{1/2} W_0 \widehat{\lambda} \\ o_p(1) &= n^{1/2} W_0' (\widehat{\phi} - \phi_0) \end{aligned}$$

and hence

$$\begin{bmatrix} n^{1/2} \frac{\partial}{\partial \phi} Q(\phi_0) \\ 0 \end{bmatrix} = \begin{bmatrix} K_0 & W_0 \\ W_0' & 0 \end{bmatrix} \begin{bmatrix} n^{1/2} (\widehat{\phi} - \phi_0) \\ n^{1/2} \widehat{\lambda} \end{bmatrix} + o_p(1).$$

If

$$n^{1/2} \frac{\partial}{\partial \phi} Q(\phi_0) \xrightarrow{d} \mathcal{N}(0, L)$$

then

$$n^{1/2} \begin{pmatrix} \widehat{\phi} - \phi_0 \\ \widehat{\lambda} \end{pmatrix} \xrightarrow{d} \mathcal{N} \left( 0, \begin{bmatrix} K_0 & W_0 \\ W_0' & 0 \end{bmatrix}^{-1} \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} K_0 & W_0 \\ W_0' & 0 \end{bmatrix}^{-1} \right)$$

provided that

$$\begin{vmatrix} K_0 & W_0 \\ W_0' & 0 \end{vmatrix} \neq 0,$$

which holds true only if there are not redundant constraints in  $w(\phi)$ . The last condition can be viewed as an identifiability condition. If  $\phi_0$  is not identifiable without  $w(\phi) = 0$  then  $K_0$  is singular and the incorporation of  $w(\phi) = 0$  leads to non-singularity (see Theorem 6). On the other hand, if  $\phi$  is identifiable without the constraints  $w(\phi) = 0$ , then  $K_0$  is already non-singular.

If  $L = K_0$  (for instance if  $Q(\cdot) = -n^{-1}$  times log likelihood), then the *Asymptotic Covariance* matrix becomes

$$\begin{bmatrix} K_0 & W_0 \\ W_0' & 0 \end{bmatrix}^{-1} \begin{bmatrix} K_0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} K_0 & W_0 \\ W_0' & 0 \end{bmatrix}^{-1}.$$

Moreover,  $K_0$  is non-singular, the last displayed expression becomes

$$\begin{bmatrix} K_0^{-1} \left( I - W_0 (W_0' K_0^{-1} W_0)^{-1} W_0' K_0^{-1} \right) & 0 \\ 0 & (W_0' K_0^{-1} W_0)^{-1} \end{bmatrix}$$

whereas the unrestricted estimator has an *Asymptotic Covariance* matrix  $K_0^{-1}$ , which is greater than or equal to

$$K_0^{-1} \left( I - W_0 (W_0' K_0^{-1} W_0)^{-1} W_0' K_0^{-1} \right)$$

because

$$K_0^{-1} W_0 (W_0' K_0^{-1} W_0)^{-1} W_0' K_0^{-1} \geq 0.$$

Thus, the constraints in  $w$  play a double role

**#1** To obtain the identifiability of the system (if needed).

**#2** To improve the efficiency of the estimators.

### 5.4.1 CONSISTENCY OF RESTRICTED LSE OF $A_0$

Assume

$$\alpha = \text{vec}(A') = F\theta - f$$

and consider the least squares, for which

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta)$$

where

$$Q(\theta) = \text{tr} \left( A \frac{X'X}{n} A' \right).$$

**Theorem 30** *Assuming A5, a sufficient condition for*

$$\hat{\alpha} \xrightarrow{P} \alpha_0 = \text{vec}(A'_0)$$

*is that*

$$F' \text{vec} \left( \begin{array}{c} B_0^{-1} \Sigma_0 \\ 0 \end{array} \right) = 0.$$

**Proof.** Using the fact that  $\text{tr}(EBCD) = \text{vec}'(C)(D \otimes B')\text{vec}(E')$ , we have that

$$\begin{aligned}
 Q(\theta) &= \text{vec}'(A') \left( I_G \otimes \frac{X'X}{n} \right) \text{vec}(A') \\
 &= \alpha' \left( I_G \otimes \frac{X'X}{n} \right) \alpha \\
 &= \alpha' \widehat{L} \alpha \\
 &= (F\theta - f)' \widehat{L} (F\theta - f),
 \end{aligned}$$

which implies that

$$\widehat{\theta} = \left( F' \widehat{L} F \right)^{-1} F' \widehat{L} f.$$

The last expression looks like the *GLS* estimator of  $f$  on  $F$  with covariance matrix  $L^{-1}$ .

So, for the consistency it suffices to examine the matrix  $F'\widehat{L}F$ . But we have that  $\widehat{L} = I_G \otimes \frac{X'X}{n}$  with

$$\begin{aligned} \frac{X'X}{n} &= \frac{1}{n} \begin{bmatrix} Y' \\ Z' \end{bmatrix} [Y, Z] \\ &= \begin{bmatrix} n^{-1}Y'Y & n^{-1}Y'Z \\ * & n^{-1}Z'Z \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \Pi_0 M \Pi'_0 + \Omega_0 & \Pi_0 M \\ * & M \end{bmatrix}. \end{aligned}$$

Thus,

$$\begin{aligned} \widehat{L} &\xrightarrow{P} \left[ I_G \otimes \begin{pmatrix} \Pi_0 M \Pi'_0 + \Omega_0 & \Pi_0 M \\ * & M \end{pmatrix} \right] \\ &= \left[ I_G \otimes \left\{ \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Pi_0 \\ I \end{pmatrix} M (\Pi'_0, I) \right\} \right] =: L. \end{aligned}$$

Hence

$$\widehat{\theta} \xrightarrow{P} \theta_0$$

if

$$\widehat{\theta} = (F'\widehat{L}F)^{-1} F'\widehat{L}f \xrightarrow{P} \theta_0,$$

that is

$$(F'LF)^{-1} F'Lf = \theta_0,$$

or similarly that

$$F'Lf = (F'LF) \theta_0.$$

So, we have obtained that a sufficient condition for the consistency is that

$$F'L(F\theta_0 - f) = 0,$$

or in terms of  $\alpha$ , that

$$F'L\alpha_0 = 0.$$

Next, using that  $vec(ABC) = (C' \otimes A) vec(B)$ , we have that the left side of the last displayed equality is

$$\begin{aligned}
& F' \left[ I_G \otimes \left\{ \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Pi_0 \\ I \end{pmatrix} M(\Pi'_0, I) \right\} \right] vec(A'_0) \\
&= F' vec \left[ \left\{ \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Pi_0 \\ I \end{pmatrix} M(\Pi'_0, I) \right\} A'_0 \right] \\
&= F' vec \left[ \begin{pmatrix} \Omega_0 B'_0 \\ 0 \end{pmatrix} + \begin{pmatrix} \Pi_0 \\ I \end{pmatrix} M \times 0 \right] \quad (\text{because } B\Pi + C = 0) \\
&= F' vec \begin{pmatrix} B_0^{-1} \Sigma_0 \\ 0 \end{pmatrix}.
\end{aligned}$$

Thus, we conclude that  $\hat{\theta} \xrightarrow{P} \theta_0$  implies that  $\hat{\alpha} \xrightarrow{P} \alpha_0$  with the right side being a sufficient condition. ■

**Example 20**  $\hat{\alpha} \xrightarrow{P} \alpha_0$  when  $B_0 = I_G$  a priori. That is, we have the classical multiple regression model.

**Example 21**  $\hat{\alpha} \xrightarrow{p} \alpha_0$  in a recursive system.

## 6 ESTIMATION OF NONLINEAR SEM AND TRANSFORMATION MODELS

Consider the complete system of equations

$$u(x_i; \theta) = u_i,$$

where  $x_i = (y'_i, z'_i)'$  and  $u_i$  is a  $G \times 1$  vector. Also all throughout, we shall assume that

1.  $u_i$  and  $z_i$  are independent.
2.  $Eu_i = 0$  for all  $i = 1, \dots, n$ .
3.  $E(u_i u'_j | z_i, z_j) = E(u_i u'_j) = \Sigma \mathcal{I}(i = j)$ .

## 6.1 MAXIMUM LIKELIHOOD ESTIMATOR (GAUSSIAN)

Let's assume that  $u_i \simeq \mathcal{N}(0, \Sigma)$  for all  $i = 1, \dots, n$ . Then, the objective function becomes

$$\begin{aligned} Q(\theta, \Sigma) &= \frac{1}{2} \log |\Sigma| - \frac{1}{n} \sum_{i=1}^n \log \left| \frac{\partial u(x_i; \theta)}{\partial y'_i} \right| \\ &\quad + \frac{1}{2n} \sum_{i=1}^n u(x_i; \theta)' \Sigma^{-1} u(x_i; \theta). \end{aligned}$$

Recall that  $\log |\partial u(x_i; \theta) / \partial y'_i|$  is nothing but the *Jacobian* of the transformation of  $u_i = u(x_i; \theta)$ .

The function  $Q(\theta, \Sigma)$  is just  $-1/2n$  times the log likelihood function.

The first derivatives of  $Q(\theta)$  are

$$\begin{aligned} \frac{\partial}{\partial \theta_j} Q(\theta, \Sigma) &= \frac{1}{n} \sum_{i=1}^n \left\{ -\text{tr} \left[ \left( \frac{\partial u(x_i; \theta)}{\partial y'_i} \right)^{-1} \frac{\partial^2 u(x_i; \theta)}{\partial \theta_j \partial y'_i} \right] \right. \\ &\quad \left. + u(x_i; \theta)' \Sigma^{-1} \frac{\partial u(x_i; \theta)}{\partial \theta_j} \right\}. \end{aligned}$$

Therefore

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} Q(\theta, \Sigma) &= \frac{1}{n} \sum_{i=1}^n \left\{ u(x_i; \theta)' \Sigma^{-1} T_j(x_i; \theta) \right. \\
&\quad \left. - \text{tr} \left[ \left( \frac{\partial u(x_i; \theta)}{\partial y'_i} \right)^{-1} \frac{\partial T_j(x_i; \theta)}{\partial y'_i} \right] \right\} \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ u(x_i; \theta)' \Sigma^{-1} T_j(x_i; \theta) - \text{tr} \left( \frac{\partial T_j(x_i; \theta)}{\partial u'_i} \right) \right\},
\end{aligned} \tag{6.1}$$

where

$$T_j(x_i; \theta) = \frac{\partial u(x_i; \theta)}{\partial \theta_j}.$$

Also, let the reduced form be

$$y_i = R(u_i, z_i; \theta_0).$$

It can be shown, see Amemiya (1977, 1982)<sup>6</sup>, that

$$E \left( \frac{\partial}{\partial \theta_j} Q(\theta_0, \Sigma_0) \right) = 0, \quad (6.2)$$

which suggests that under regularity conditions

$$\left( \widehat{\theta}, \widehat{\Sigma} \right) = \arg \min_{\theta, \Sigma} Q(\theta, \Sigma)$$

satisfies that  $\widehat{\theta} \rightarrow_P \theta_0$ .

Moreover, under some additional regularity conditions

$$n^{1/2} \left( \widehat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, \Xi^{-1} \right),$$

where, as usual,

$$\frac{\partial^2 Q(\theta_0, \Sigma_0)}{\partial \theta \partial \theta'} \xrightarrow{P} \Xi$$

---

<sup>6</sup>Recall that in the linear SEM,

$$\frac{\partial Q(\theta_0)}{\partial \theta} = - \begin{bmatrix} P' & W' \end{bmatrix} \begin{bmatrix} \Omega^{-1} \otimes I_K & 0 \\ 0 & \frac{1}{2} (\Omega^{-1} \otimes \Omega^{-1}) \end{bmatrix} \begin{bmatrix} \text{vec}(\widehat{N}) \\ \text{vec}(\widehat{O} - \Omega_0) \end{bmatrix}$$

which implies that  $\hat{\theta}$  is asymptotically efficient.

In the linear case, we have seen that these types of results hold true under very general class of distributions for the error  $u$ . Basically all that it was needed for the consistency was that  $E(u|z) = 0$ .

## 6.2 INSTRUMENTAL VARIABLES. GMM ESTIMATES

These type of estimators will allow us to obtain consistent estimators of  $\theta_0$  under wide range of distributional assumptions. The basic idea is to use the fact that if  $u_i$  and  $z_i$  are independent (implying that the conditional expectation is zero) then we have that

$$\begin{aligned}\forall \phi(z_i); \quad E[u_i \phi(z_i)] &= E\{E[u_i \phi(z_i) | z_i]\} \\ &= E\{\phi(z_i) E[u_i | z_i]\} \\ &= 0.\end{aligned}\tag{4}$$

Thus, one can expect that

$$\frac{1}{n} \sum_{i=1}^n \phi(z_i) u_i \xrightarrow{P} 0.$$

Thus, consider

$$\tilde{\theta} = \arg \min_{\theta \in \Theta} Q(\theta),$$

where

$$Q(\theta) = \frac{1}{n} \sum_{i=1}^n u(x_i; \theta)' P_i' \left( \sum_{i=1}^n P_i M P_i' \right)^{-1} \sum_{i=1}^n P_i u(x_i; \theta), \quad (5)$$

and where  $P_i$  is a  $p \times G$  matrix of functions of  $z_i$  and  $M$  is a positive definite matrix.

Compare with the classical *IVE* in the linear regression case, what you have there is

$$\begin{aligned} & \frac{1}{n} (Y - Z\beta)' W (W'W)^{-1} W' (Y - Z\beta) \\ &= \frac{1}{n} \sum_{i=1}^n u_i w_i' \left( \sum_{i=1}^n w_i w_i' \right)^{-1} \sum_{i=1}^n w_i u_i. \end{aligned}$$

The first order conditions are

$$\frac{1}{n} (Z'W) (W'W)^{-1} W' (Y - Z\tilde{\beta}) = 0$$

Because  $|Z'W| \neq 0$  and  $|W'W| \neq 0$ , then the previous equation becomes

$$W' (Y - Z\tilde{\beta}) = 0 \Rightarrow \tilde{\beta} = (W'Z)^{-1} W'Y.$$

Now, the first order conditions for (5) are

$$\frac{\partial Q(\theta)}{\partial \theta} = \frac{2}{n} \sum_{i=1}^n T(x_i; \theta)' P_i' \left( \sum_{i=1}^n P_i M P_i' \right)^{-1} \sum_{i=1}^n P_i u(x_i; \theta),$$

where as before

$$T(x_i; \theta) = \frac{\partial}{\partial \theta'} u(x_i; \theta).$$

But, because  $z_i$  and  $u_i$  are independent, we can expect that

$$\begin{aligned} E(P_i u(x_i; \theta_0)) &= 0 \\ &= E P_i E u(x_i; \theta_0). \end{aligned}$$

Then,

$$\frac{1}{n} \sum_{i=1}^n P_i u(x_i; \theta_0) \xrightarrow{P} 0.$$

Also, we can expect that

$$\frac{1}{n} \sum_{i=1}^n T(x_i; \theta_0)' P_i' \xrightarrow{P} \mathcal{A},$$

where  $\mathcal{A}$  is a finite matrix, and also that

$$\frac{1}{n} \sum_{i=1}^n P_i M P_i' \xrightarrow{P} \Xi > 0.$$

So, gathering the three last previous results, they would suggest that

$$\tilde{\theta} \xrightarrow{P} \theta_0. \tag{6}$$

This is what, among other issues, Hansen (1982) showed.

**Remark 11** *Observe that we have not made any assumption regarding the probability density function of  $u_i$ . Only we have assumed that  $E(u_i) = 0$  and thus it appears that (6) holds true under very general conditions.*

*Later on we will see who is the matrix  $\sum_{i=1}^n P_i M P_i'$  and which instruments  $P_i$  to choose to obtain efficiency.*

*In fact, we can show that, under suitable regularity conditions,*

$$n^{1/2} \left( \tilde{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, \mathcal{A}^{-1} \mathcal{B} \mathcal{A}^{-1} \right),$$

where

$$\begin{aligned}\mathcal{A} &= p \lim \frac{1}{n} \sum_{i=1}^n T'(x_i; \theta_0) P_i' \\ &= p \lim \frac{1}{n} \sum_{i=1}^n S'(z_i; \theta_0) P_i' \\ \mathcal{B} &= p \lim \frac{1}{n} \sum_{i=1}^n P_i' \Sigma_0 P_i,\end{aligned}$$

with

$$S(z; \theta) = E[T(x; \theta) | z].$$

Let

$$\mathcal{C} = p \lim \frac{1}{n} \sum_{i=1}^n S'(z_i; \theta_0) \Sigma_0^{-1} S(z_i; \theta_0),$$

then we can show that

$$\mathcal{C}^{-1} \leq \mathcal{A}^{-1} \mathcal{B} \mathcal{A}^{-1}$$

and therefore  $\mathcal{C}^{-1}$  is a lower bound, which can be achieved if we choose

$$P_i = \Sigma_0^{-1} S(z_i; \theta_0); \quad M = \Sigma_0.$$

That is, if

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n u(x_i; \theta)' \Sigma_0^{-1} S_i \left( \sum_{i=1}^n S_i' \Sigma_0^{-1} S_i \right)^{-1} \sum_{i=1}^n S_i' \Sigma_0^{-1} u(x_i; \theta),$$

then  $\hat{\theta}$  is efficient, with  $S_i = S(z_i; \theta_0)$ .

**Problem:**

$\Sigma_0$  and  $S_i$  are unknown. So, the above estimator  $\hat{\theta}$  is infeasible (it cannot be computed).

With regard to  $\Sigma_0$ , there are no problems because we can employ  $\tilde{\theta}$  to obtain  $\hat{u}_i, i = 1, \dots, n$ , and then from here  $\hat{\Sigma}$  as usual.

But what about  $S_i$ ? Recall that by definition,

$$\begin{aligned} S_i &= E [T(x_i; \theta_0) | z_i] \\ &= E \left[ \frac{\partial}{\partial \theta'} u(x_i; \theta_0) | z_i \right] \end{aligned}$$

which is?

Because the partial derivative is not linear in  $u_i$ , this latter conditional expectation is difficult to know, or put it in another way, unless we know something about the distribution of  $u_i$  is not possible to obtain  $S_i$ . However, if we know something else about  $u_i$  apart from  $E(u_i) = 0$  and  $E(u_i u_j') = \Sigma \mathcal{I}(i = j)$ , then  $\hat{\theta}$  would not be efficient. Chamberlain (1987, 1992).

### 6.2.1 FEASIBLE OPTIMAL IV ESTIMATOR

Recall

$$\begin{aligned}u(x_i; \theta) &= u_i; & T(x_i; \theta) &= \frac{\partial}{\partial \theta'} u(x_i; \theta); \\S(z_i; \theta) &= E[T(x_i; \theta) | z_i].\end{aligned}$$

Assume that there exists a vector  $v_i \in \mathbb{R}^r$  independent of  $z_i$  and  $\xi_0 \in \mathbb{R}^\ell$  such that

$$T(x_i; \theta_0) = Q(v_i, z_i; \xi_0), \quad i = 1, \dots, n.$$

Now, because  $v_1$  and  $z_1$  are independent

$$E[Q(v_1, z_1; \xi_0) | z_1] = E[T(x_1; \theta_0) | z_1] = S_1.$$

Therefore, this leads us to think how we can estimate  $S_i$ ,  $i = 1, \dots, n$ .

Given observations  $v_1, v_2, \dots, v_n$ , the typical estimator is as usual to replace

population moments by sample moments, that is

$$\begin{aligned}\widehat{S}_i &= \frac{1}{M_i} \sum_{j \in \mathcal{M}_i} Q_{ji} \\ &= \frac{1}{M_i} \sum_{j \in \mathcal{M}_i} Q(v_j, z_i; \xi_0),\end{aligned}$$

where  $\mathcal{M}_i$  consists of  $M_i$  of the indices  $1, \dots, n$ , and where  $v_j$ ,  $j \in \mathcal{M}_i$ , are independent of  $z_i$ .

Because  $v_i$  and  $\xi_0$  are not observable, all we need to do is to replace them by say  $\widetilde{v}_i$  and  $\widehat{\xi}_0$ , respectively, which are “consistent” for  $v_i$  and  $\xi_0$ .

In the original paper, Robinson (1988) discusses three leading situations. We will only focus on one of them. (Case I in his paper.)

Let

$$y = \mathcal{R}(u, z; \theta_0)$$

be (the reduced form) a unique solution for

$$u(x; \theta_0) = u.$$

In this case

$$\begin{aligned} Q(v, z; \xi) &= Q(u, z; \theta) \\ &= T(\mathcal{R}(u, z; \theta_0), z; \theta) \end{aligned}$$

and choose  $v_i \equiv u_i$  and  $\xi \equiv \theta$ . The idea is

$$\begin{aligned} T(x; \theta) &= T(y, z; \theta) \\ &= T(\mathcal{R}(u, z; \theta_0), z; \theta) \\ &= Q(u, z; \theta) \end{aligned}$$

implies that  $u = v$  and  $\xi_0 = \theta_0$ .

Obviously both  $U$  and  $\theta_0$  are unknown and thus we will write instead of  $u, \tilde{u}$ , e.g. obtained by a previous estimator of  $\theta$  and  $\theta$  by such an estimator.

In this case,

$$\begin{aligned} S(z_i; \theta) &= E[T(x_i; \theta) | z_i] \\ &= E[Q(u_i, z_i; \theta) | z_i], \end{aligned}$$

which will be estimated by

$$\bar{S}_i = \frac{1}{M_i} \sum_{j \in \mathcal{M}_i} Q(\tilde{u}_j, z_i; \tilde{\theta}).$$

Notice that there would be a close form for  $Q$  if there was one for “ $\mathcal{R}$ ”. Otherwise, we need to use numerical approximation. One possible reason to allow  $\mathcal{M}_i$  to be a subset of  $\{1, \dots, n\}$  is for numerical or computational savings. Thus, the feasible *IVE* will be

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n u(x_i; \theta)' \hat{\Sigma}^{-1} \bar{S}_i \left( \sum_{i=1}^n \bar{S}_i' \hat{\Sigma}^{-1} \bar{S}_i \right)^{-1} \sum_{i=1}^n \bar{S}_i' \hat{\Sigma}^{-1} u(x_i; \theta),$$

which under regularity conditions

$$n^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{C}^{-1})$$

and a consistent estimator for  $\mathcal{C}$  is given by

$$\frac{1}{n} \sum_{i=1}^n \bar{S}_i' \hat{\Sigma}^{-1} \bar{S}_i \xrightarrow{P} \mathcal{C}.$$

$\mathcal{M}_i$  can increase arbitrary slow as  $n \rightarrow \infty$ , in fact all we need is that  $\{\mathcal{M}_i\}_{i=1}^n$  satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{M}_i^{-1} = 0.$$

This will save a lot of computations. In general we will need  $\sum_{i=1}^n \mathcal{M}_i$  operations for the estimation of  $\theta_0$ . (When  $P_i$  is used, we only need  $n$  operations.)

### **WHEN $u_i$ 's ARE SERIALY CORRELATED**

In this situation, the previous estimator is not efficient. (Compare with the *GLS* estimator in linear regression models.)

Obviously, we assume that there are not lags of  $y_i$  in the model. Now, put

$$U(\theta) = (u(x_1; \theta)', \dots, u(x_n; \theta)')$$

and

$$\Omega_0 = E[U(\theta_0)U'(\theta_0)].$$

Then the optimal *IVE* is in this case

$$\widehat{\theta}^* = \arg \min_{\theta \in \Theta} U'(\theta) \Omega_0^{-1} S (S' \Omega_0^{-1} S)^{-1} S' \Omega_0^{-1} U(\theta),$$

where  $S = (S_1, \dots, S_n)'$  and

$$n^{1/2} (\widehat{\theta}^* - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Psi^{-1}),$$

with

$$\Psi = p \lim_{n \rightarrow \infty} S' \Omega_0^{-1} S.$$

Let's look at one example.

**Example 22** *Assume a particular parametric specification for  $\Omega$ . For instance, suppose that the sequence  $\{u_i\}_{i \in \mathbb{Z}}$  admits an  $AR(\infty)$  representation*

$$\sum_{j=0}^{\infty} B_j u_{i-j} = e_i; \quad \sum_{j=0}^{\infty} |B_j| < \infty,$$

where  $\{e_i\}_{i \in \mathbb{Z}}$  is a white noise sequence of random variables, that is  $Ee_i = 0$  and  $E(e_i e_j') = I_{G \times G} \mathcal{I}(i = j)$ .

Because the model is parametric, it means that  $B_j = B_j(\tau_0)$  for all  $j \geq 0$ .  
 Now, suppose that we have some estimates of  $\tau_0$ , say  $\hat{\tau}$ , such that

$$\hat{\tau} - \tau_0 = O_p\left(n^{-1/2}\right).$$

Then it is pretty obvious that we can estimate  $B_j$  by  $\hat{B}_j = B_j(\hat{\tau})$ .

Then, apply the feasible transformation

$$\tilde{e}_i(\theta) = \sum_{j=0}^{i-1} \hat{B}_j u(x_{i-j}; \theta),$$

and the instruments will be not  $\bar{S}_i$ , but

$$\hat{H}_i = \sum_{j=0}^{i-1} \hat{B}_j \bar{S}_{i-j}$$

with the feasible estimator of  $\theta_0$  given by

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{i=1}^n \tilde{e}_i'(\theta) \hat{H}_i \left( \sum_{i=1}^n \hat{H}_i' \hat{H}_i \right)^{-1} \sum_{i=1}^n \hat{H}_i' \tilde{e}_i(\theta).$$

*Under suitable conditions, we can show that*

$$n^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Psi^{-1}),$$

*where*

$$\hat{\Psi} = \frac{1}{n} \sum_{i=1}^n \hat{H}_i' \hat{H}_i \xrightarrow{P} \Psi.$$

## 6.2.2 Linear SEM

We are going to see that the *2SLS* or the *3SLS* are instrumental variable estimators.

In short, they are like setting the variable  $P_i$  in GMM as

$$P_i = I_G \otimes z_i, \text{ or } P_i = (I_G \otimes z_i) \Sigma^{-1/2}.$$

Recall that the objective function of 2SLS is

$$u_1' P_Z u_1 + \cdots + u_G' P_Z u_G.$$

First recall that our model is

$$AX' = U'$$

which after using the *vec* notation becomes

$$(X \otimes I_G) \alpha = u,$$

where  $\alpha = \text{vec}(A)$  and  $u = \text{vec}(U')$ .

Now, let  $W$  be a  $nG \times (G(G + K))$  matrix and consider

$$\hat{\theta}; \quad \hat{\alpha} = \alpha(\hat{\theta})$$

such that

$$\begin{aligned} Q(\hat{\theta}) &= \min_{\theta} Q(\theta) \\ Q(\theta) &= \alpha'W'(X \otimes I_G)\alpha, \end{aligned}$$

where  $W$  is a matrix of instruments. Then,  $\hat{\theta}$  and  $\hat{\alpha}$  are the *IVE* of  $\theta_0$  and  $\alpha_0$  respectively.

**Theorem 31** *The 2SLS is an IVE where*

$$W = Z \begin{pmatrix} \tilde{\Pi} \\ I \end{pmatrix}' \otimes I_G$$

and  $\tilde{\Pi}$  is the LSE of  $\Pi_0$ , that is

$$\begin{aligned} Y &= Z\Pi' + V \\ \tilde{\Pi} &= Y'Z(Z'Z)^{-1}. \end{aligned}$$

**Proof.** First, we know that the objective function for the 2SLS is

$$\begin{aligned} Q(\theta) &= \text{tr} \left\{ AX'Z(Z'Z)^{-1}Z'XA' \right\} \\ &= \text{vec}'(A) \text{vec} \left( AX'Z(Z'Z)^{-1}Z'X \right) \\ &= \alpha' \left( X'Z(Z'Z)^{-1}Z'X \otimes I_G \right) \text{vec}(A) \\ &= \alpha' \left( X'Z(Z'Z)^{-1}Z'X \otimes I_G \right) \alpha. \end{aligned}$$

Now,

$$\begin{aligned} X'Z(Z'Z)^{-1}Z'X \otimes I_G &= \left( X'Z(Z'Z)^{-1}Z' \otimes I_G \right) (X \otimes I_G) \\ &= \left[ \begin{pmatrix} Y'Z(Z'Z)^{-1} \\ (Z'Z)(Z'Z)^{-1} \end{pmatrix} Z' \otimes I_G \right] (X \otimes I_G) \\ &= W'(X \otimes I_G) \end{aligned}$$

with

$$W' = \left[ \left( \begin{array}{c} \tilde{\Pi} \\ I \end{array} \right) Z' \otimes I_G \right].$$

Therefore,

$$Q(\theta) = \alpha' W' (X \otimes I_G) \alpha$$

which concludes the proof. ■

**Remark 12** *For the 3SLS we have that*

$$W = \left[ Z \left( \begin{array}{c} \tilde{\Pi} \\ I \end{array} \right)' \otimes \hat{\Sigma}^{-1} \right].$$

## 7 HYPOTHESIS TESTING

- We will focus on tests with large sample justification. They are based on the properties of point estimates that we have already examined and described.
- We split the parameter vector  $\theta$  as

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix},$$

where  $\theta_1$  is  $q \times 1$  and  $\theta_2$  is a  $s \times 1$  with  $p = q + s$ .

- The true value is  $\theta_0 = (\theta'_{10}, \theta'_{20})'$  and the type of hypothesis that we will look at are

$$\begin{aligned} H_0 &: \theta_{10} = 0 \begin{cases} \text{composite if } q < p \\ \text{simple if } q = p \end{cases} \\ H_1 &: \theta_{10} \neq 0. \end{aligned}$$

**Remark 13** *We shall consider always  $\theta_{10} = 0$  without loss of generality.*

**Example 23** *Suppose*

$$H_0 : R_{q \times p} \phi_0 = r$$

*with rank*  $(R) = q$ . *Then,*

$$\theta_1 = R\phi - r$$

$$\theta_2 = \phi_2,$$

*so that*

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} R & \\ 0 & I_{p-q} \end{pmatrix} \phi - \begin{pmatrix} r \\ 0 \end{pmatrix}$$

*which implies that*

$$\phi = \begin{pmatrix} R & \\ 0 & I_{p-q} \end{pmatrix}^{-1} \left( \theta + \begin{pmatrix} r \\ 0 \end{pmatrix} \right).$$

## 7.1 Review of Classical Hypothesis Testing

Let a partition of the parameter space  $\Theta$  be made and given by

$$\Theta = \Theta_0 \cup \Theta_1$$

The *null hypothesis* is given by

$$\mathcal{H}_0 : \theta \in \Theta_0$$

and is tested against the *alternative hypothesis*

$$\mathcal{H}_1 : \theta \in \Theta_1$$

The null hypothesis  $H_0$  is maintained unless it is rejected in favor of the alternative hypothesis  $H_1$ .

- When  $\Theta_0$  and  $\Theta_1$  are singleton sets, we say that the hypothesis is *simple*. Otherwise, they are *composite*.

- A ‘test’ is thus completely synonymous to a ‘critical region’. We will therefore refer to a test with its critical region. That is, the state space  $\mathcal{X}$  is partitioned as the disjoint union of the *critical region*  $C$  and *acceptance region*  $A$ , i.e.,

$$\mathcal{X} = C \cup A$$

If  $x \in C$ , then  $H_0$  is rejected in favor of  $H_1$ . If, on the other hand,  $x \in A$ , then  $H_0$  is continued to be maintained.

- The statistical hypothesis testing is usually based on a *test statistic*  $\tau$ .

Let  $X = (X_1, \dots, X_n)'$  be a random sample, and suppose the distribution of  $X$  is given by a parametric family  $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ .

- The *power function*  $\pi(\theta)$  of the test  $C$  is defined by

$$\pi(\theta) = P_\theta(C)$$

Moreover,

$$\max_{\theta \in \Theta_0} \pi(\theta)$$

is called the *size* of the test, while the values of  $\pi$  at  $\theta \in \Theta_1$  are called the *power* of the test.

We define

**Definition 18** *The test  $C^*$  is Uniformly Most Powerful (UMP) if  $\forall \theta \in \Theta_1$*

$$P_\theta(C^*) \geq P_\theta(C)$$

*for any test represented by  $C$  of the same size.*

- When both the null and alternative hypotheses are simple, we may write  $H_0 = \theta_0$  and  $H_1 = \theta_1$ . Since  $\Theta = \{\theta_0, \theta_1\}$  in this case,  $\mathcal{P}$  consists of two distributions, which we write as

$$P_{\theta_0} = P_0 \quad \text{and} \quad P_{\theta_1} = P_1$$

for the null and alternative distributions, respectively. Clearly,  $P_0(C)$  and  $P_1(C)$  are the size and power of the test.

- Notice that  $P_0(C)$  is the probability of rejecting  $H_0$  when it is true. On the other hand,  $P_1(A)$  is the probability of accepting  $H_0$  when  $H_0$  is false (and  $H_1$  is true). Both of  $P_0(C)$  and  $P_1(A)$  are the probabilities of making errors, which we refer to as the *type I* and *type II* errors, respectively.

Assume that both the null and alternative hypotheses are simple, and the distributions  $P_0$  and  $P_1$  are given by the likelihood functions  $p(x, \theta_0)$  and  $p(x, \theta_1)$ .

**Lemma 32 (Neyman-Pearson)** *The test which rejects  $\mathcal{H}_0$  when*

$$\lambda(x) = \frac{p(x, \theta_1)}{p(x, \theta_0)} \geq c \text{ or } \lambda(x) = \frac{p(x, \theta_0)}{p(x, \theta_1)} \leq c$$

*for a constant  $c$  is most powerful. In words, for any size  $\alpha$ , the likelihood ratio critical region is the best critical region.*

**Proof** Let  $C^* = \{x \mid \lambda(x) \geq c\}$ , and suppose  $C$  is any test other than  $C^*$  with the same size, i.e.  $P_0(C) = P_0(C^*)$ . We need to show  $P_1(C) \leq P_1(C^*)$ . Assume without loss of generality that  $C$  and  $C^*$  are disjoint. Then

$$p(x, \theta_1) \geq c p(x, \theta_0) \quad \text{over } C^*$$

and

$$p(x, \theta_1) < c p(x, \theta_0) \quad \text{over } C$$

We thus have

$$P_1(C^*) = \int_{C^*} p(x, \theta_1) dx \geq c \int_{C^*} p(x, \theta_0) dx = c P_0(C^*)$$

and

$$P_1(C) = \int_C p(x, \theta_1) dx < c \int_C p(x, \theta_0) dx = c P_0(C).$$

The stated result then follows immediately from the fact that  $c P_0(C^*) = c P_0(C)$ .

■

- The above lemma guarantees the existence of a best test under simple null and alternative hypotheses. The criterion used in the construction of the likelihood ratio (LR) test  $\lambda(x) \geq c$  indeed tells us about the form of the partition of  $\mathcal{X}$ . We can view  $\lambda(x)$  as a ratio of marginal power to marginal size. Then the LR test includes only those points in  $\mathcal{X}$  that have significant enough power increase per unit of size increase.
- The LR test is generalized as

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_1} p(x, \theta)}{\sup_{\theta \in \Theta_0} p(x, \theta)}$$

for composite hypotheses. The generalized LR test rejects  $\mathcal{H}_0$  when  $\tau(x) \geq c$ , where  $\tau(x)$  is any monotone increasing function of  $\lambda(x)$  and  $c$  is given for a prescribed size. Note that the Neyman-Pearson lemma does not apply to the generalized LR test. Optimality properties of the generalized LR tests are much harder to show.

- In practice,  $c$  should be chosen properly to control the size of the test. It is called ‘critical value’ of the test.

- The regression model is a semi-parametric model, in which the parameter of interest  $\beta$  is parametric while the distribution family is nonparametric in the sense that it depends on infinite-dimensional unknowns. In this setting, the size cannot be calculated **exactly** but is calculated **asymptotically**. Here comes the “asymptotic test.”
- The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing. An alternative approach, associate with Fisher, is to report an asymptotic p-value. The asymptotic p-value for a statistic is constructed as follows. Let a statistic  $t_n$  converge in distribution to  $Z$ . Define the tail probability, or asymptotic p-value function

$$p(t) = \Pr \{|Z| \geq |t|\}$$

Then the asymptotic p-value of the statistic  $t_n$  is

$$p_n = p(t_n).$$

Sometimes the asymptotic p-value function is defined as

$$p(t) = \Pr\{Z \geq t\}.$$

- Another helpful observation is that the p-value function has simply made a unit-free transformation of the test statistic. That is, under the null,  $p_n \rightarrow^d U[0, 1]$ , so the “unusualness” of the test statistic can be compared to the easy-to-understand uniform distribution, regardless of the complication of the distribution of the original test statistic.
- In applications, our model is often given in a form other than likelihood function, as in the linear regression. Even in the MLE framework, other test statistics than LR statistic are employed in practice from various reasons. The t/Wald test and score (LM) test are most common and constitute the trinity of tests including the LR test.

## 7.2 WALD TEST (Generalized)

Assume that

**H1**  $\widehat{\theta}$  is a consistent estimator of  $\theta_0 = (\theta'_{10}, \theta'_{20})'$  such that

$$n^{1/2} (\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{D})$$

for all  $\theta_0 \in \Theta$  and where

$$\mathcal{D} = \mathcal{D}(\theta_0) = \begin{pmatrix} \mathcal{D}_{11}(\theta_0) & \mathcal{D}_{12}(\theta_0) \\ * & \mathcal{D}_{22}(\theta_0) \end{pmatrix}$$

and  $\mathcal{D}_{11}(\theta_0)$  and  $\mathcal{D}_{22}(\theta_0)$  are  $q \times q$  and  $p - q \times p - q$  respectively.

**H2** Under  $H_0 : \theta_{10} = 0$

$$\mathcal{D}_{11}(\widehat{\theta}) \xrightarrow{P} \mathcal{D}_{11} \begin{pmatrix} 0 \\ \theta_{20} \end{pmatrix} = \mathcal{D}_{11}^0.$$

**Definition 19** (Generalized) Wald statistic is given by

$$\mathcal{W} = n\widehat{\theta}'_1\widehat{\mathcal{D}}_{11}^{-1}\widehat{\theta}_1.$$

**Remark 14** Many possible estimators of  $\theta_0$ ,  $\widehat{\theta}$ , exist in a given problem and for a given  $\widehat{\theta}$  there are many choices of  $\widehat{\mathcal{D}}_{11}$ .  $\widehat{\mathcal{D}}_{11}$  may or may not make use of  $H_0$ . However, if it does, it will depend on  $\widehat{\theta}_2$  only. If not, we have that  $\widehat{\mathcal{D}}_{11} \rightarrow_p \mathcal{D}_{11}(\theta_0)$  whether or not  $H_0$  holds true. If we employ the MLE, then  $\mathcal{D}^{-1}$  is the information matrix.

**Example 24** Consider the linear regression model

$$y_i = \beta_0' z_i + v_i,$$

$Ev_i = 0$ ,  $Ev_i v_j = \sigma_v^2 \mathcal{I}(i = j)$ . Now our hypothesis testing is

$$H_0 : R\beta_0 = r$$

and  $\text{rank}(R) = q$ . Write

$$\theta_1 = R\beta - r$$

$$\theta_2 = \beta_2$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Let  $\widehat{\beta}$  be the LSE, so that

$$\widehat{\theta} = \begin{pmatrix} R\widehat{\beta} - r \\ \widehat{\beta}_2 \end{pmatrix}.$$

Assume

$$\begin{aligned}\widehat{M} &= \frac{1}{n} Z' Z \xrightarrow{P} M > 0 \\ n^{1/2} (\widehat{\beta} - \beta) &\xrightarrow{d} \mathcal{N}(0, \sigma_v^2 M^{-1}).\end{aligned}$$

Therefore,

$$n^{1/2} R (\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma_v^2 R M^{-1} R').$$

Now under  $H_0$  :

$$\begin{aligned}n^{1/2} \widehat{\theta}_1 &= n^{1/2} (R \widehat{\beta} - r) \\ &= n^{1/2} R (\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{11})\end{aligned}$$

where  $\mathcal{D}_{11} = \sigma_v^2 R M^{-1} R'$ .

If

$$\widehat{\sigma}_v^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \widehat{\beta}' z_i)^2,$$

then

$$\mathcal{W} = n \widehat{\theta}_1' \left( \widehat{\sigma}_v^2 R \widehat{M}^{-1} R' \right)^{-1} \widehat{\theta}_1.$$

**Theorem 33** *Assume H1 and H2. Then, under  $H_0 : \theta_{10} = 0$ ,*

$$\mathcal{W} \xrightarrow{d} \chi_q^2.$$

**Proof.**  $H_0$  and  $H1$  imply that

$$n^{1/2}\widehat{\theta}_1 = n^{1/2} \left( \widehat{\theta}_1 - \theta_{10} \right) \xrightarrow{d} \mathcal{N} \left( 0, \mathcal{D}_{11} \right)$$

so that

$$n\widehat{\theta}'_1 \mathcal{D}_{11}^{-1} \widehat{\theta}_1 \xrightarrow{d} \chi_q^2$$

by continuous mapping theorem.

But,

$$\begin{aligned} \mathcal{W} &= n\widehat{\theta}'_1 \widehat{\mathcal{D}}_{11}^{-1} \widehat{\theta}_1 = n\widehat{\theta}'_1 \mathcal{D}_{11}^{-1} \widehat{\theta}_1 + n^{1/2} \widehat{\theta}'_1 \left( \widehat{\mathcal{D}}_{11}^{-1} - \mathcal{D}_{11}^{-1} \right) \widehat{\theta}_1 n^{1/2} \\ &= n\widehat{\theta}'_1 \mathcal{D}_{11}^{-1} \widehat{\theta}_1 + O_p(1) o_p(1) O_p(1) \\ &= n\widehat{\theta}'_1 \mathcal{D}_{11}^{-1} \widehat{\theta}_1 + o_p(1) \\ &\xrightarrow{d} \chi_q^2. \end{aligned}$$

■

Thus, this theorem is telling us that an asymptotic  $\alpha$ -level test is given by

$$\text{Reject } H_0 \text{ iff } W > \chi_{q,\alpha}^2,$$

where

$$\int_{\chi_{q,\alpha}^2}^{\infty} pdf(\chi_q^2) d\chi_q^2 = \alpha.$$

**Definition 20** For a test-statistic  $\hat{\tau}$  suppose we reject  $H_0$  when  $\hat{\tau} > c$ . Then, the function defined as

$$\Pi^c(\theta_{10}) = \Pr\{\hat{\tau} > c | \theta_{10}\}$$

is called the POWER FUNCTION for  $\hat{\tau}$ .

**Theorem 34** Under  $H1$  and  $H2$ , for  $c = \chi_{q,\alpha}^2$ , we obtain that

$$\Pi^c(0) \xrightarrow[n \rightarrow \infty]{} \alpha.$$

**Proof.** Trivial. ■

**Definition 21** The test-statistic  $\hat{\tau}$  is consistent IFF

$$\Pi^c(\theta_{10}) \xrightarrow[n \rightarrow \infty]{} 1$$

for all  $\theta_{10} \neq 0$  and for all  $c > 0$ .

What does it mean? Basically that you will reject the null hypothesis  $H_0 : \theta_{10} = 0$ , with probability ONE as  $n \rightarrow \infty$  if  $H_0$  false.

Let's introduce the additional assumption.

**H3** For all  $\theta_0 \in \Theta$ ,

$$\widehat{\mathcal{D}}_{11} \xrightarrow{P} \mathcal{D}_{11} > 0.$$

**Theorem 35** *Assuming H1 and B1, the Wald test is consistent for all  $\alpha$ .*

**Proof.** By H1,

$$n^{1/2} \left( \widehat{\theta}_1 - \theta_{10} \right) \xrightarrow{d} \mathcal{N} \left( 0, \mathcal{D}_{11} \left( \theta_{10} \right) \right).$$

Therefore,

$$n^{1/2} \widehat{\theta}_1 = n^{1/2} \theta_{10} + O_p(1).$$

Next, H3 implies that

$$\begin{aligned} \mathcal{W} &= n^{1/2} \widehat{\theta}'_1 \widehat{\mathcal{D}}_{11}^{-1} \widehat{\theta}_1 n^{1/2} \\ &= \left( n^{1/2} \theta'_{10} + O_p(1) \right) \left( \mathcal{D}_{11}^{-1} + o_p(1) \right) \left( n^{1/2} \theta_{10} + O_p(1) \right) \\ &= n \theta'_{10} \mathcal{D}_{11}^{-1} \theta_{10} + O_p \left( n^{1/2} \right). \end{aligned}$$

Then,

$$\begin{aligned}\Pi^c(\theta_{10}) &= \Pr\{\mathcal{W} > c|\theta_{10}\} \\ &= \Pr\left\{n\theta'_{10}\mathcal{D}_{11}^{-1}\theta_{10} + O_p\left(n^{1/2}\right) > c|\theta_{10}\right\} \\ &\xrightarrow[n \rightarrow \infty]{} 1\end{aligned}$$

because

$$n\theta'_{10}\mathcal{D}_{11}^{-1}\theta_{10} > 0$$

and dominates  $O_p\left(n^{1/2}\right)$  as  $n \uparrow \infty$ . This is true for all  $\alpha$  and for all  $c > 0$ . ■

**Remark 15** *As it was mentioned above,  $\widehat{\mathcal{D}}_{11}$  may converge to a different value under  $H_1$  if the estimator is based on  $H_0$ , whereas an estimate  $\widetilde{\mathcal{D}}_{11}$  which it is not based on  $H_0$  satisfies*

$$\widetilde{\mathcal{D}}_{11} \xrightarrow{P} \mathcal{D}_{11}(\theta_{10})$$

*for all  $\theta_0$ .*

The next question is the following. For a given problem, one may have more than one possible consistent estimator of  $\theta_0$  and all of them may converge in distribution to a normal distribution. Thus, the Wald tests based on these different estimators will have a (asymptotic)  $\chi_q^2$ -distribution. In addition, all of them are *consistent*. Thus, which one shall we choose?

For that purpose, we will consider the distribution of the test under *LOCAL-ALTERNATIVES* or *PITMAN-ALTERNATIVES*.

- Consider

$$H_{1n} \equiv \theta_{10}^n = \delta n^{-1/2}$$

for a fixed  $q \times 1$  vector  $\delta$ . The choice of  $\delta$  determines the *direction* of departure from the null  $H_0$ . For instance if  $\delta = (1, 0, \dots)'$ , then it means that the departure is in the direction of the first coordinate of  $\theta$ . In a sense, we do not consider a model but a sequence of models indexed by the sample size “ $n$ ”.

**H4** Under  $H_{1n}$ ,

$$\begin{aligned}n^{1/2} \left( \widehat{\theta}_1 - \theta_{10}^n \right) &= n^{1/2} \widehat{\theta}_1 - \delta \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{11}) \\ \widehat{\mathcal{D}}_{11} \xrightarrow{P} \mathcal{D}_{11} &> 0,\end{aligned}$$

where  $\mathcal{D}_{11}$  is the same as in Assumption *H2*.

**Definition 22** *The random variable  $X$  follows a non-central  $\chi_q^2$  distribution with non-centrality parameter*

$$\Lambda = \sum_{j=1}^q \lambda_j^2$$

*( $X \simeq \chi_q^2(\Lambda)$ ) if*

$$X = \sum_{j=1}^q (u_j + \lambda_j)^2,$$

*where  $\{u_j\}_{j=1}^q$  are iid $\mathcal{N}(0, 1)$  random variables.*

CDF functions of noncentral chisquare distribution

We now give a theorem without proof.

**Theorem 36** *For fixed  $c$  and  $q$ , the function*

$$\Pr \{ \chi_q^2(\Lambda) > c \}$$

*is increasing in  $\Lambda$ .*

**Theorem 37** Under B2,

$$\mathcal{W} \xrightarrow{d} \chi_q^2 (\delta' \mathcal{D}_{11}^{-1} \delta).$$

**Proof.**  $H_4$  implies that

$$n^{1/2} \widehat{\theta}_1 \xrightarrow{d} \mathcal{N}(\delta, \mathcal{D}_{11}).$$

Thus,

$$\begin{aligned} n^{1/2} \mathcal{D}_{11}^{-1/2} \widehat{\theta}_1 &\xrightarrow{d} \mathcal{N}(\mathcal{D}_{11}^{-1/2} \delta, I) \\ &\equiv \mathcal{N}(\lambda, I), \end{aligned}$$

where  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_q)' = \mathcal{D}_{11}^{-1/2} \delta$ .

So,

$$a_n =: n^{1/2} \mathcal{D}_{11}^{-1/2} \widehat{\theta}_1 - \lambda \xrightarrow{d} \mathcal{N}(0, I)$$

and hence we can conclude that  $a_n$  is a  $q$ -dimensional vector of random variables whose joint distribution behaves like  $iid\mathcal{N}(0, 1)$ . Now, the test is

$$\mathcal{W} = n \widehat{\theta}_1' \widehat{\mathcal{D}}_{11}^{-1} \widehat{\theta}_1 = n^{1/2} \widehat{\theta}_1' \mathcal{D}_{11}^{-1/2} \mathcal{D}_{11}^{-1/2} \widehat{\theta}_1 n^{1/2} + o_p(1)$$

because

$$\widehat{\mathcal{D}}_{11} \xrightarrow{P} \mathcal{D}_{11} > 0.$$

But,

$$\begin{aligned} n^{1/2} \widehat{\theta}'_1 \mathcal{D}_{11}^{-1/2} \mathcal{D}_{11}^{-1/2} \widehat{\theta}_1 n^{1/2} &= (a_n + \lambda)' (a_n + \lambda) \\ &= \sum_{i=1}^q (a_{ni} + \lambda_i)^2 \stackrel{d}{=} \chi_q^2 (\lambda' \lambda) \\ &\equiv \chi_q^2 (\delta' \mathcal{D}_{11}^{-1} \delta). \end{aligned}$$

This concludes the proof. ■

**Definition 23** Consider two statistics  $\mathcal{W}_1$  and  $\mathcal{W}_2$ , where

$$\mathcal{W}_i \xrightarrow{d} \chi_q^2$$

under  $H_0$  for  $i = 1, 2$  and

$$\mathcal{W}_i \xrightarrow{d} \chi_q^2(\Lambda_i)$$

under  $H_{1n}$ . Then,

We say that  $\mathcal{W}_1$  is as efficient as  $\mathcal{W}_2$  if  $\Lambda_1 = \Lambda_2$ .

In the same way, we say that  $\mathcal{W}_1$  is more efficient than  $\mathcal{W}_2$  if  $\Lambda_1 - \Lambda_2 > 0$ .

What the previous definition says is that

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pi_1 \left( \delta n^{-1/2} \right) &= \Pr \{ \chi_q^2(\Lambda_1) > c \} \\ &> \Pr \{ \chi_q^2(\Lambda_2) > c \} = \lim_{n \rightarrow \infty} \Pi_2 \left( \delta n^{-1/2} \right), \end{aligned}$$

where the inequality comes from the last theorem.

**Theorem 38** Let  $\hat{\theta}$  and  $\tilde{\theta}$  be two consistent estimators with Asymptotic Covariance Matrix given by  $\mathcal{D}$  and  $\mathcal{E}$ , respectively. Then, the condition  $\mathcal{D}_{11} < \mathcal{E}_{11}$

implies that the Wald test based on  $\widehat{\theta}$  is more efficient than the Wald test based on  $\widetilde{\theta}$ .

**Proof.** Let

$$\begin{aligned}\mathcal{W}_1 &= n\widetilde{\theta}'_1\widehat{\mathcal{D}}_{11}^{-1}\widehat{\theta}_1 \\ \mathcal{W}_2 &= n\widetilde{\theta}'_1\widehat{\mathcal{E}}_{11}^{-1}\widetilde{\theta}_1,\end{aligned}$$

where  $\widehat{\mathcal{D}}_{11}$  and  $\widehat{\mathcal{E}}_{11}$  are consistent estimators of  $\mathcal{D}_{11}$  and  $\mathcal{E}_{11}$ , respectively. Then, from Theorem 67, we know that

$$\begin{aligned}\mathcal{W}_1 &\xrightarrow{d} \chi_q^2(\delta'\mathcal{D}_{11}^{-1}\delta) \\ \mathcal{W}_2 &\xrightarrow{d} \chi_q^2(\delta'\mathcal{E}_{11}^{-1}\delta).\end{aligned}$$

So,

$$\begin{aligned}\Lambda_1 &= \delta'\mathcal{D}_{11}^{-1}\delta \\ \Lambda_2 &= \delta'\mathcal{E}_{11}^{-1}\delta\end{aligned}$$

which implies from Theorem 68 that

$$\Lambda_1 - \Lambda_2 = \delta' (\mathcal{D}_{11}^{-1} - \mathcal{E}_{11}^{-1}) \delta > 0$$

because  $\mathcal{E}_{11} - \mathcal{D}_{11} > 0$ . ■

The last theorem tell us that the test has more or bigger power the more precise (efficient) the estimator of  $\theta_0$  is.

**Example 25** Consider our linear regression model in Example 48. We wish to test  $H_0 : R\beta - r = 0$ .

We have two possible estimators, namely

$$\begin{aligned}\hat{\beta}_{LSE} &= (Z'Z)^{-1} Z'Y \\ \hat{\beta}_{IVE} &= (S'Z)^{-1} S'Y,\end{aligned}$$

where  $S$  is a set of instruments. Then, we have that

$$\hat{\theta}_{LSE} = R\hat{\beta}_{LSE} - r; \quad \hat{\theta}_{IVE} = R\hat{\beta}_{IVE} - r$$

and the corresponding Wald tests are respectively

$$\begin{aligned}\mathcal{W}_1 &= n\hat{\theta}_{LSE} \left( \hat{\sigma}_{LSE}^2 R \left( \frac{Z'Z}{n} \right)^{-1} R' \right)^{-1} \hat{\theta}_{LSE} \\ \mathcal{W}_2 &= n\hat{\theta}_{IVE} \left( \hat{\sigma}_{IVE}^2 R \left\{ \left( \frac{S'Z}{n} \right)^{-1} \left( \frac{S'S}{n} \right) \left( \frac{Z'S}{n} \right)^{-1} \right\} R' \right)^{-1} \hat{\theta}_{IVE}.\end{aligned}$$

We already know that

$$Z'S(S'S)^{-1}S'Z \leq Z'Z.$$

Also, in this case we know that

$$\begin{aligned}\Lambda_1 &= \delta' \left( \sigma^2 R \text{plim} \left( \frac{Z'Z}{n} \right)^{-1} R' \right)^{-1} \delta \\ \Lambda_2 &= \delta' \left( \sigma^2 R \text{plim} \left\{ \left( \frac{S'Z}{n} \right)^{-1} \left( \frac{S'S}{n} \right) \left( \frac{Z'S}{n} \right)^{-1} \right\} R' \right)^{-1} \delta.\end{aligned}$$

Thus,  $\mathcal{W}_1$  would be more efficient than  $\mathcal{W}_2$  if

$$\left( R \text{plim} \left( \frac{Z'Z}{n} \right)^{-1} R' \right)^{-1} \geq \left( R \text{plim} \left\{ \left( \frac{S'Z}{n} \right)^{-1} \left( \frac{S'S}{n} \right) \left( \frac{Z'S}{n} \right)^{-1} \right\} R' \right)^{-1}$$

or equivalently if

$$R \text{plim} \left( \frac{Z'Z}{n} \right)^{-1} R' \leq R \text{plim} \left\{ \left( \frac{S'Z}{n} \right)^{-1} \left( \frac{S'S}{n} \right) \left( \frac{Z'S}{n} \right)^{-1} \right\} R'.$$

Because  $R$  is a full rank matrix, the latter inequality holds true iff

$$\text{plim} \left( \frac{Z'Z}{n} \right)^{-1} \leq \text{plim} \left\{ \left( \frac{S'Z}{n} \right)^{-1} \left( \frac{S'S}{n} \right) \left( \frac{Z'S}{n} \right)^{-1} \right\}$$

or

$$plim \left\{ \left( \frac{Z'S}{n} \right) \left( \frac{S'S}{n} \right)^{-1} \left( \frac{S'Z}{n} \right) \right\} \leq plim \left( \frac{Z'Z}{n} \right)$$

which is the case because

$$Z'S (S'S)^{-1} S'Z \leq Z'Z.$$

This concludes.

### 7.2.1 CONSIDER EXTREMUM ESTIMATES

As usual, our estimator of  $\theta_0$  is given by

$$\widehat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta).$$

Write

$$Q_{\theta}(\theta) = \frac{\partial}{\partial \theta} Q(\theta); \quad Q_{\theta\theta}(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} Q(\theta),$$

where

$$\begin{aligned} n^{1/2} Q_{\theta}(\theta_0) &\xrightarrow{d} \mathcal{N}(0, \mathcal{B}) \\ Q_{\theta\theta}(\theta_0) &\xrightarrow{P} \mathcal{A}, \end{aligned}$$

and  $\mathcal{A} > 0$ .

Next, we already know that under suitable regularity conditions,

$$n^{1/2} (\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{A}^{-1} \mathcal{B} \mathcal{A}^{-1})$$

and we have efficiency if  $\mathcal{A} = \mathcal{B}$ , in which case

$$n^{1/2} \left( \widehat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, \mathcal{A}^{-1} \right).$$

Cases where  $\mathcal{A} = \mathcal{B}$  holds true are the *PMLE*, or *MDE* or the *3SLSE* in the linear *Simultaneous Equation System*

$$A(\theta) x_i = u_i,$$

when  $\{u_i\}_{i \in \mathbb{Z}}$  is a sequence of uncorrelated homoscedastic sequence,  $\Sigma = E(u_i u_i')$  is unrestricted and  $\theta_0$  only parameterizes  $\mathcal{A}$ .

Assume that  $\mathcal{A} = \mathcal{B}$ , that is

**H5**

$$n^{1/2}Q_\theta(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}^{-1})$$

and if  $\bar{\theta} \rightarrow_P \theta_0$ , then for all  $\theta_0 \in \Theta$

$$Q_{\theta\theta}(\bar{\theta}) \xrightarrow{P} \mathcal{D}^{-1} > 0.$$

The Wald statistic is defined as

$$\mathcal{W} = n\hat{\theta}'_1 \hat{\mathcal{D}}_{11}^{-1} \hat{\theta}_1.$$

**Theorem 39** *Assuming that H1, H2, and H5 hold, then under  $H_0$ ,*

$$\mathcal{W} \xrightarrow{d} \chi_q^2.$$

**Proof.** By definition,

$$Q_\theta(\hat{\theta}) = 0 = Q_\theta(\theta_0) + Q_{\theta\theta}(\bar{\theta})(\hat{\theta} - \theta_0),$$

where  $\bar{\theta}$  is an intermediate point between  $\hat{\theta}$  and  $\theta_0$ .

Now by *H5*,

$$Q_{\theta\theta}(\bar{\theta}) \xrightarrow{P} \mathcal{D}^{-1}$$

so,

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}),$$

and hence

$$n^{1/2}(\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{11}).$$

The latter implies that

$$\begin{aligned} \mathcal{W} &= n\hat{\theta}'_1 \hat{\mathcal{D}}_{11}^{-1} \hat{\theta}_1 \\ &= n\hat{\theta}'_1 \mathcal{D}_{11}^{-1} \hat{\theta}_1 + n^{1/2} \hat{\theta}'_1 (\hat{\mathcal{D}}_{11}^{-1} - \mathcal{D}_{11}^{-1}) n^{1/2} \hat{\theta}_1 \\ &= n\hat{\theta}'_1 \mathcal{D}_{11}^{-1} \hat{\theta}_1 + O_p(1) o_p(1) O_p(1) \\ &\xrightarrow{d} \chi_q^2 \end{aligned}$$

which completes the proof. ■

**Theorem 40** *Assuming that B1 and H5 hold,  $\mathcal{W}$  is a consistent test.*

**Proof.** We already know that

$$n^{1/2}\widehat{\theta}_1 = n^{1/2}\theta_{10} + O_p(1).$$

Thus,

$$\left(n^{1/2}\widehat{\theta}_1\right)^{-1} \xrightarrow{P} 0,$$

which implies that

$$\mathcal{W}^{-1} \xrightarrow{P} 0$$

and hence that for all  $c > 0$ ,

$$\Pr\{\mathcal{W} > c\} \rightarrow 1.$$

■

Consider the local alternatives

$$H_{1n} : n^{-1/2}\delta.$$

**H6** Under  $H_{1n}$ ,

$$\widehat{\mathcal{D}}_{11} \xrightarrow{P} \mathcal{D}_{11}.$$

**Theorem 41** *Assume H5 and H6. Then, under  $H_{1n}$  we have that*

$$\mathcal{W} \xrightarrow{d} \chi_q^2 (\delta' \mathcal{D}_{11}^{-1} \delta).$$

**Proof.** By definition,

$$Q_\theta (\widehat{\theta}) = 0 = Q_\theta (\theta_0) + Q_{\theta\theta} (\bar{\theta}) (\widehat{\theta} - \theta_0).$$

Thus, by A5,

$$\begin{aligned} n^{1/2} (\widehat{\theta} - \theta_0) &= -Q_{\theta\theta}^{-1} (\bar{\theta}) Q_\theta (\theta_0) \\ &\xrightarrow{d} \mathcal{N} (0, \mathcal{D}) \end{aligned}$$

which implies that

$$n^{1/2} (\widehat{\theta}_1 - \theta_{10}) \xrightarrow{d} \mathcal{N} (0, \mathcal{D}_{11}).$$

hence

$$n^{1/2}\widehat{\theta}_1 - \delta \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{11}) \equiv n^{1/2}\widehat{\theta}_1 \xrightarrow{d} \mathcal{N}(\delta, \mathcal{D}_{11}).$$

So, we can conclude by the continuous mapping theorem that

$$\begin{aligned} \mathcal{W} &= n^{1/2}\widehat{\theta}'_1 \widehat{\mathcal{D}}_{11}^{-1} n^{1/2}\widehat{\theta}_1 \\ &\xrightarrow{d} \chi_q^2(\delta' \mathcal{D}_{11}^{-1} \delta). \end{aligned}$$

■

### 7.3 THE LANGRANGE MULTIPLIER TEST (SCORE TEST)

- The Wald test is based on the estimator of  $\theta_0$  under the alternative, and then it checks whether  $\hat{\theta}$  agrees or not with  $H_0$ . Sometimes it is easier to estimate the model under  $H_0$ , and the *Langragean Multiplier* ( $\mathcal{LM}$ ) test makes use of this. Let

$$\tilde{\theta} = \arg \min_{\theta \in \Theta: \theta_1=0} Q(\theta).$$

That is,  $\tilde{\theta} = (0', \tilde{\theta}_2)'$ . By definition,

$$Q_2(\tilde{\theta}) = 0,$$

where

$$Q_i = \frac{\partial}{\partial \theta_i} Q(\theta), \quad i = 1, 2.$$

- The standard way to obtain  $\tilde{\theta}$  is through the lagrangean objective function

$$\mathcal{L}(\theta; \lambda) = Q(\theta) - \lambda' \theta_1.$$

The first order conditions are

$$\begin{aligned}\frac{\partial}{\partial \theta_1} \mathcal{L}(\tilde{\theta}; \tilde{\lambda}) &= Q_1(\tilde{\theta}) - \tilde{\lambda} = 0 \\ \frac{\partial}{\partial \theta_2} \mathcal{L}(\tilde{\theta}; \tilde{\lambda}) &= Q_2(\tilde{\theta}) = 0.\end{aligned}$$

From the first of the last two displayed equations, we obtain that

$$\tilde{\lambda} = Q_1(\tilde{\theta})$$

whereas from the second one we obtain  $\tilde{\theta}$  by solving the system.

- The intuition of the test is that if  $H_0$  were true then the constraints employed to optimize  $Q(\theta)$  would be superfluous. Equivalently, does  $\tilde{\theta}$  approximately satisfy the conditions for a minimum? Is  $Q_1(\tilde{\theta}) = 0$ ?

**H7** Under  $H_0$ ,

$$\tilde{\mathcal{D}}_{11} \xrightarrow{P} \mathcal{D}_{11} \begin{pmatrix} 0 \\ \theta_{20} \end{pmatrix}.$$

**Definition 24** The  $\mathcal{LM}$  statistic (Lagrangean Multiplier) is defined as

$$\mathcal{LM} = n\tilde{\lambda}'\tilde{\mathcal{D}}_{11}\tilde{\lambda}.$$

**Remark 16** Because

$$Q_2(\tilde{\theta}) = 0$$

we have that the  $\mathcal{LM}$  statistic can be written as

$$\mathcal{LM} = nQ'_\theta(\tilde{\theta})\tilde{\mathcal{D}}Q_\theta(\tilde{\theta}),$$

and this is why the test is sometimes called the SCORE test.

**Example 26** (*Cont. regression model*)

$$H_0 : R\beta_0 - r = 0$$

where  $\text{rank}(R) = q$  and  $\theta_{10} = R\beta - r$  and  $\theta_{20} = (\beta_2', \sigma^2)'$ . So,

$$\begin{pmatrix} \theta_{10} \\ \theta_{20} \end{pmatrix} = \begin{pmatrix} R & 0 \\ 0_{(p-q) \times q} & I_{(p-q) \times (p-q)} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix} - \begin{pmatrix} r \\ 0 \\ 0 \end{pmatrix}$$

or equivalently

$$\begin{aligned} \theta &= \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}; \quad S = \begin{pmatrix} R & \\ 0_{(p-q) \times q} & I_{(p-q) \times (p-q)} \end{pmatrix} \\ s &= \begin{pmatrix} r \\ 0 \end{pmatrix}. \end{aligned}$$

(Now  $\sigma^2$  has been removed.)

If we remove  $\sigma^2$ , we have that

$$\theta_* = S\beta - s \implies \beta = S^{-1}(\theta_* + s).$$

Then, the PMLE will be

$$\begin{aligned} Q(\theta) &= \frac{1}{2} \log \sigma^2 + \frac{1}{2n\sigma^2} (Y - Z\beta)' (Y - Z\beta) \\ &= \frac{1}{2} \log \sigma^2 + \frac{1}{2n\sigma^2} (Y - ZS^{-1}(\theta_* + s))' (Y - ZS^{-1}(\theta_* + s)). \end{aligned}$$

The first order conditions are

$$\begin{aligned} Q_{\theta_*}(\tilde{\theta}) &= -\frac{1}{n\tilde{\sigma}^2} (S^{-1}Z') (Y - ZS^{-1}(\tilde{\theta}_* + s)) \\ Q_{\sigma^2}(\tilde{\theta}) &= \frac{1}{\tilde{\sigma}^2} - \frac{1}{n\tilde{\sigma}^4} (Y - ZS^{-1}(\tilde{\theta}_* + s))' (Y - ZS^{-1}(\tilde{\theta}_* + s)). \end{aligned}$$

So, we have that

$$\tilde{\sigma}^2 = \frac{1}{n} (Y - Z\tilde{\beta})' (Y - Z\tilde{\beta}).$$

Also

$$Q_{\theta_*\theta_*}(\bar{\theta}) = \frac{1}{\tilde{\sigma}^2} \left( S^{-1} \frac{Z'Z}{n} S'^{-1} \right) \xrightarrow{P} \mathcal{D}^{-1}.$$

Therefore,

$$\begin{aligned}\mathcal{LM} &= nQ_{\theta_*'}(\tilde{\theta})Q_{\theta_*\theta_*}^{-1}(\bar{\theta})Q_{\theta_*}(\tilde{\theta}) \\ &= (Y - Z\tilde{\beta})' \frac{Z(Z'Z)^{-1}Z'}{\tilde{\sigma}^2} (Y - Z\tilde{\beta}) \\ &= nR^2,\end{aligned}$$

because  $Y - Z\tilde{\beta}$  are the residuals least squares. That is, the LM is  $n$  times the coefficient of multiple correlation in the regression of  $\tilde{u}_i$  on  $z_i$ . That is, if the least squares estimate in that regression model is denoted by  $\tilde{\alpha}$ , then

$$R^2 = \frac{\tilde{\alpha}'Z'Z\tilde{\alpha}}{\tilde{u}'\tilde{u}}.$$

**Theorem 42** *Assuming H5 and H7, under  $H_0$ , we have that*

$$\mathcal{LM} \xrightarrow{d} \chi_q^2.$$

**Proof.** Recall that H5 indicates that  $n^{1/2}Q_\theta(\theta_0) \rightarrow_d \mathcal{N}(0, \mathcal{D}^{-1})$  and H7 that  $\tilde{\mathcal{D}}_{11} \rightarrow_P \mathcal{D}_{11}(0, \theta'_{20})$ .

Now, put  $Q_i^0 = Q_i(0, \theta'_{20})$  with  $Q_\theta^0 = (Q_1^0, Q_2^0)'$ . Then, by the Mean Value Theorem, we obtain that

$$\begin{aligned}\tilde{\lambda} &= Q_1(\tilde{\theta}) = Q_1^0 + \bar{Q}_{12}(\tilde{\theta}_2 - \theta_{20}) \\ 0 &= Q_2(\tilde{\theta}) = Q_2^0 + \bar{Q}_{22}(\tilde{\theta}_2 - \theta_{20}).\end{aligned}$$

(Recall that  $\theta_{10} = 0$ , so that it does not change as is a fixed quantity.)

Hence,

$$\tilde{\theta}_2 - \theta_{20} = -\bar{Q}_{22}^{-1} Q_2^0$$

which implies that

$$\begin{aligned}
 n^{1/2}\tilde{\lambda} &= n^{1/2} \left( I; -\bar{Q}_{12}\bar{Q}_{22}^{-1} \right) \begin{pmatrix} Q_1^0 \\ Q_2^0 \end{pmatrix} \\
 &= \left( I; -\bar{Q}_{12}\bar{Q}_{22}^{-1} \right) n^{1/2} Q_\theta^0 \\
 &\xrightarrow{d} \mathcal{N}(0, \mathcal{F}),
 \end{aligned}$$

where

$$\begin{aligned}
 \mathcal{F} &= \left( I; -Q_{12}^0 Q_{22}^{0-1} \right) \begin{pmatrix} \mathcal{D}^{11} & \mathcal{D}^{12} \\ \mathcal{D}^{21} & \mathcal{D}^{22} \end{pmatrix} \begin{pmatrix} I \\ -Q_{22}^{0-1} Q_{21}^0 \end{pmatrix} \\
 &= \left( \mathcal{D}^{11} - \mathcal{D}^{12} (\mathcal{D}^{22})^{-1} \mathcal{D}^{21} \right) \\
 &= \mathcal{D}_{11}^{-1}.
 \end{aligned}$$

Then, we can conclude that

$$\begin{aligned}\mathcal{LM} &= n\tilde{\lambda}'\widehat{\mathcal{D}}_{11}\tilde{\lambda} \\ &= n\tilde{\lambda}'\mathcal{D}_{11}\tilde{\lambda} + n^{1/2}\tilde{\lambda}'\left(\widehat{\mathcal{D}}_{11} - \mathcal{D}_{11}\right)n^{1/2}\tilde{\lambda} \\ &\xrightarrow{d}\chi_q^2\end{aligned}$$

because  $n^{1/2}\tilde{\lambda} \rightarrow_d \mathcal{N}(0, \mathcal{D}_{11}^{-1})$ .

■

**Theorem 43** *Under H5 and B1 (e.g.  $\tilde{\mathcal{D}}_{11} \rightarrow_P \mathcal{D}_{11}$  for all  $\theta \in \Theta$ ), we have that the  $\mathcal{LM}$  test is consistent.*

**Proof.** We already know that

$$\begin{aligned} n^{1/2}\tilde{\lambda} &= n^{1/2} \left( I; -\bar{Q}_{12}\bar{Q}_{22}^{-1} \right) \begin{pmatrix} Q_1^0 \\ Q_2^0 \end{pmatrix} \\ &= \left( I; -\bar{Q}_{12}\bar{Q}_{22}^{-1} \right) n^{1/2} \left( Q_\theta \begin{pmatrix} \theta_{10} \\ \theta_{20} \end{pmatrix} - \frac{\partial}{\partial \theta'_1} Q_\theta \begin{pmatrix} \tilde{\theta}_1 \\ \theta_{20} \end{pmatrix} \theta_{10} \right) \end{aligned}$$

by the mean value theorem. So, we have that

$$n^{1/2}\tilde{\lambda} = O_p(1) + n^{1/2} (Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}) \theta_{10},$$

which implies that

$$n^{-1/2}\tilde{\lambda}^{-1} \xrightarrow{P} 0$$

and hence the consistency of the test.

■

**Theorem 44** *Assume H6. Then, under  $H_{1n}$ , we have that*

$$\mathcal{LM} \xrightarrow{d} \chi_q^2(\delta' \mathcal{D}_{11}^{-1} \delta).$$

**Proof.** Proceeding as with the proof of the previous theorem, we have that

$$\begin{aligned} n^{1/2} \tilde{\lambda} &= \left( I; -\bar{Q}_{12} \bar{Q}_{22}^{-1} \right) n^{1/2} \left( Q_\theta \begin{pmatrix} \theta_{10} \\ \theta_{20} \end{pmatrix} - \frac{\partial}{\partial \theta'_1} Q_\theta \begin{pmatrix} \tilde{\theta}_1 \\ \theta_{20} \end{pmatrix} n^{-1/2} \delta \right) \\ &\xrightarrow{d} \mathcal{N}(-\mathcal{D}_{11}^{-1} \delta, \mathcal{D}_{11}^{-1}). \end{aligned}$$

Therefore,

$$\mathcal{D}_{11}^{1/2} n^{1/2} \tilde{\lambda} \xrightarrow{d} \mathcal{N}(-\mathcal{D}_{11}^{-1/2} \delta, I)$$

which implies that

$$\mathcal{LM} \xrightarrow{d} \chi_q^2(\delta' \mathcal{D}_{11}^{-1} \delta).$$

■

**Remark 17** *Theorems 73 and 75 show that the Wald and the Lagrangian Multiplier tests are asymptotically equivalent.*

## 7.4 THE PSEUDO LIKELIHOOD RATIO TEST

We shall define the *Pseudo Likelihood Ratio* ( $\mathcal{LR}$ ) test as follows.

**Definition 25** *The (pseudo) likelihood ratio test is given by*

$$\mathcal{LR} = 2n \left( Q \left( \tilde{\theta} \right) - Q \left( \hat{\theta} \right) \right).$$

**Example 27** (*Cont. regression model*) Again our null hypothesis  $H_0$  is

$$H_0 : R\beta - r = 0.$$

On the other hand,

$$Q(\theta) = \frac{1}{2} \log \sigma^2 + \frac{1}{2n} \frac{(Y - Z\beta)'(Y - Z\beta)}{\sigma^2}.$$

So, we have that

$$\begin{aligned} Q(\hat{\theta}) &= \frac{1}{2} \log \hat{\sigma}^2 + \frac{1}{2n} \frac{(Y - Z\hat{\beta})'(Y - Z\hat{\beta})}{\hat{\sigma}^2} \\ &= \frac{1}{2} \log \hat{\sigma}^2 + \frac{1}{2}, \end{aligned}$$

because

$$\hat{\sigma}^2 = \frac{1}{n} (Y - Z\hat{\beta})'(Y - Z\hat{\beta}).$$

Similarly,

$$\begin{aligned} Q(\tilde{\theta}) &= \frac{1}{2} \log \tilde{\sigma}^2 + \frac{1}{2n} \frac{(Y - Z\tilde{\beta})'(Y - Z\tilde{\beta})}{\tilde{\sigma}^2} \\ &= \frac{1}{2} \log \tilde{\sigma}^2 + \frac{1}{2}, \end{aligned}$$

where  $\tilde{\beta}$  is the restricted least squares estimator of  $\beta$ .

Hence, we conclude that in this case

$$\mathcal{LR} = n \log \left( \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right).$$

Now,

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n} (Y - Z\tilde{\beta})'(Y - Z\tilde{\beta}) \\ &= \hat{\sigma}^2 + (\hat{\beta} - \tilde{\beta})' \hat{M} (\hat{\beta} - \tilde{\beta}) \end{aligned}$$

because  $Z$  is orthogonal to  $\hat{U}$ .

On the other hand, the restricted least squares estimator  $\tilde{\beta}$  is

$$\tilde{\beta} = \hat{\beta} + \widehat{M}^{-1}R' \left( R\widehat{M}^{-1}R' \right)^{-1} \left( r - R\hat{\beta} \right).$$

Then, we obtain that

$$\begin{aligned} \tilde{\sigma}^2 &= \hat{\sigma}^2 + \left( R\hat{\beta} - r \right)' \left( R\widehat{M}^{-1}R' \right)^{-1} \left( R\hat{\beta} - r \right) \\ &= \hat{\sigma}^2 + \frac{\mathcal{W}}{n} \hat{\sigma}^2 \end{aligned}$$

or

$$\tilde{\sigma}^2 = \left( 1 + \frac{\mathcal{W}}{n} \right) \hat{\sigma}^2.$$

Then,

$$\begin{aligned} \mathcal{LR} &= n \log \left( \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right) \\ &= n \log \left( 1 + \frac{\mathcal{W}}{n} \right), \end{aligned}$$

that is, the  $\mathcal{LR}$  is a monotonic function of  $\mathcal{W}$ .

*Asymptotically, they are the same because*

$$\log\left(1 + \frac{\mathcal{W}}{n}\right) \simeq \frac{\mathcal{W}}{n} - \frac{1}{2} \frac{\mathcal{W}^2}{n^2}$$

*so that*

$$\begin{aligned}\mathcal{LR} &\simeq \mathcal{W} - \frac{1}{2} \frac{\mathcal{W}^2}{n} \\ &\equiv \mathcal{W} + o_p(1).\end{aligned}$$

**Theorem 45** *Assuming H1, H5, under  $H_0$ ,*

$$\mathcal{LR} \xrightarrow{d} \chi_q^2.$$

**Proof.** By definition,

$$\begin{aligned}\mathcal{LR} &= 2n \left( Q(\tilde{\theta}) - Q(\hat{\theta}) \right) \\ &= 2n \left( Q_{\theta}(\hat{\theta})' (\tilde{\theta} - \hat{\theta}) + \frac{1}{2} (\tilde{\theta} - \hat{\theta})' Q_{\theta\theta}(\bar{\theta}) (\tilde{\theta} - \hat{\theta}) \right) \\ &= n (\tilde{\theta} - \hat{\theta})' Q_{\theta\theta}(\bar{\theta}) (\tilde{\theta} - \hat{\theta})\end{aligned}$$

because by definition of  $\hat{\theta}$ ,  $Q_{\theta}(\hat{\theta}) = 0$  and where  $\bar{\theta}$  is an intermediate point between  $\tilde{\theta}$  and  $\hat{\theta}$ .

Now,  $Q_{\theta\theta}(\bar{\theta}) \rightarrow_P \mathcal{D}^{-1}$ . On the other hand, what about  $\tilde{\theta} - \hat{\theta}$ ?

$$0 = Q_1(\hat{\theta}) = Q_1^0 + \bar{Q}_{11}\hat{\theta}_1 + \bar{Q}_{12}(\hat{\theta}_2 - \theta_{20}) \quad (7.1)$$

$$0 = Q_2(\hat{\theta}) = Q_2^0 + \bar{Q}_{21}\hat{\theta}_1 + \bar{Q}_{22}(\hat{\theta}_2 - \theta_{20}) \quad (7.2)$$

$$0 = Q_2(\tilde{\theta}) = Q_2^0 + \check{Q}_{22}(\tilde{\theta}_2 - \theta_{20}). \quad (7.3)$$

Now, the difference between (7.2) and (7.3) leads to

$$\begin{aligned} \bar{Q}_{21}\hat{\theta}_1 + \bar{Q}_{22}(\hat{\theta}_2 - \theta_{20}) &= \check{Q}_{22}(\tilde{\theta}_2 - \theta_{20}) \\ \bar{Q}_{21}\hat{\theta}_1 + \bar{Q}_{22}(\hat{\theta}_2 - \tilde{\theta}_2) + \bar{Q}_{22}(\tilde{\theta}_2 - \theta_{20}) &= \check{Q}_{22}(\tilde{\theta}_2 - \theta_{20}) \end{aligned}$$

which implies that

$$\bar{Q}_{21}\hat{\theta}_1 + \bar{Q}_{22}(\hat{\theta}_2 - \tilde{\theta}_2) = (\check{Q}_{22} - \bar{Q}_{22})(\tilde{\theta}_2 - \theta_{20})$$

and hence that

$$\begin{aligned} n^{1/2}(\hat{\theta}_2 - \tilde{\theta}_2) &= \bar{Q}_{22}^{-1} \left\{ \bar{Q}_{21}n^{1/2}\hat{\theta}_1 + (\check{Q}_{22} - \bar{Q}_{22})n^{1/2}(\tilde{\theta}_2 - \theta_{20}) \right\} \\ &= \bar{Q}_{22}^{-1}\bar{Q}_{21}n^{1/2}\hat{\theta}_1 + o_p(1). \end{aligned}$$

So,

$$\begin{aligned}
 n^{1/2} \left( \tilde{\theta} - \hat{\theta} \right) &= \begin{pmatrix} -n^{1/2} \hat{\theta}_1 \\ n^{1/2} \left( \tilde{\theta}_2 - \hat{\theta}_2 \right) \end{pmatrix} \\
 &= \begin{pmatrix} -I \\ \overline{Q}_{22}^{-1} \overline{Q}_{21} \end{pmatrix} n^{1/2} \hat{\theta}_1 + o_p(1) \\
 &= \begin{pmatrix} -I \\ (\mathcal{D}^{22})^{-1} \mathcal{D}^{21} \end{pmatrix} n^{1/2} \hat{\theta}_1 + o_p(1)
 \end{aligned}$$

Because under  $H_0$ , we know that

$$n^{1/2} \hat{\theta}_1 \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{11}),$$

we can conclude that

$$\begin{aligned}
 \mathcal{LR} &= 2n \left( Q \left( \tilde{\theta} \right) - Q \left( \hat{\theta} \right) \right) \\
 &= n \left( \tilde{\theta} - \hat{\theta} \right)' \mathcal{D}^{-1} \left( \tilde{\theta} - \hat{\theta} \right) + o_p(1).
 \end{aligned}$$

But, also we know that

$$\left(-I; \mathcal{D}^{12} (\mathcal{D}^{22})^{-1}\right) \begin{bmatrix} \mathcal{D}^{11} & \mathcal{D}^{12} \\ \mathcal{D}^{21} & \mathcal{D}^{22} \end{bmatrix} \begin{pmatrix} -I \\ (\mathcal{D}^{22})^{-1} \mathcal{D}^{21} \end{pmatrix} = \mathcal{D}_{11}^{-1}.$$

So, we obtain that

$$\mathcal{LR} = n^{1/2} \widehat{\theta}'_1 \mathcal{D}_{11}^{-1} n^{1/2} \widehat{\theta}_1 + o_p(1) \xrightarrow{d} \chi_q^2,$$

which concludes the proof.

■

**Theorem 46** *Assuming H5, the  $\mathcal{LR}$  provides a consistent test.*

**Proof.** We know that

$$n^{1/2} (\tilde{\theta} - \hat{\theta}) = \begin{pmatrix} -I \\ \overline{Q_{22}^{-1}} \overline{Q_{21}} \end{pmatrix} \left( n^{1/2} (\hat{\theta}_1 - \theta_{10}) + n^{1/2} \theta_{10} \right).$$

From here it is standard to show that  $\mathcal{LR}^{-1} \rightarrow_P 0$  and hence the consistency of the test.

■

**Theorem 47** *Assuming H6, under  $H_{1n}$ , we have that*

$$\mathcal{LR} \xrightarrow{d} \chi_q^2 (\delta' \mathcal{D}_{11}^{-1} \delta).$$

**Proof.** As in the previous theorem, we have that

$$\begin{aligned} n^{1/2} (\tilde{\theta} - \hat{\theta}) &= \begin{pmatrix} -I \\ \overline{Q}_{22}^{-1} \overline{Q}_{21} \end{pmatrix} \left( n^{1/2} (\hat{\theta}_1 - \theta_{10}) + \delta \right) \\ &\xrightarrow{d} \mathcal{N} (\mathcal{D}_{11} \delta, \mathcal{D}_{11}^{-1}). \end{aligned}$$

From here, the remainder of the proof is standard. ■

**Remark 18** *The later theorem shows that the  $\mathcal{LR}$  is also asymptotically equivalent to the Wald and Lagrangean Multiplier test.*

**Example 28** Consider the linear regression model

$$y_i = \beta_0' z_i + v_i,$$

$Ev_i = 0$ ,  $Ev_i v_j = \sigma_v^2 \mathcal{I}(i = j)$ . Now our hypothesis testing is

$$H_0 : R\beta_0 = r$$

and  $\text{rank}(R) = q$ . Write

$$\theta_1 = R\beta - r$$

$$\theta_2 = \beta_2$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Let  $\widehat{\beta}$  be the LSE, so that

$$\widehat{\theta} = \begin{pmatrix} R\widehat{\beta} - r \\ \widehat{\beta}_2 \end{pmatrix}.$$

Assume

$$\begin{aligned}\widehat{M} &= \frac{1}{n} Z' Z \xrightarrow{P} M > 0 \\ n^{1/2} (\widehat{\beta} - \beta) &\xrightarrow{d} \mathcal{N}(0, \sigma_v^2 M^{-1}).\end{aligned}$$

Therefore,

$$n^{1/2} R (\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma_v^2 R M^{-1} R').$$

Now under  $H_0$  :

$$\begin{aligned}n^{1/2} \widehat{\theta}_1 &= n^{1/2} (R \widehat{\beta} - r) \\ &= n^{1/2} R (\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{11})\end{aligned}$$

where  $\mathcal{D}_{11} = \sigma_v^2 R M^{-1} R'$ .

If

$$\widehat{\sigma}_v^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \widehat{\beta}' z_i)^2,$$

then

$$\mathcal{W} = n \widehat{\theta}_1' \left( \widehat{\sigma}_v^2 R \widehat{M}^{-1} R' \right)^{-1} \widehat{\theta}_1.$$

**Example 29**

$$H_0 : R\beta_0 - r = 0$$

where  $\text{rank}(R) = q$  and  $\theta_{10} = R\beta - r$  and  $\theta_{20} = (\beta'_2, \sigma^2)'$ . So,

$$\begin{pmatrix} \theta_{10} \\ \theta_{20} \end{pmatrix} = \begin{pmatrix} R & 0 \\ 0_{(p-q) \times q} & I_{(p-q) \times (p-q)} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix} - \begin{pmatrix} r \\ 0 \\ 0 \end{pmatrix}$$

or equivalently

$$\begin{aligned} \theta &= \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}; \quad S = \begin{pmatrix} R & \\ 0_{(p-q) \times q} & I_{(p-q) \times (p-q)} \end{pmatrix} \\ s &= \begin{pmatrix} r \\ 0 \end{pmatrix}. \end{aligned}$$

(Now  $\sigma^2$  has been removed.)

If we remove  $\sigma^2$ , we have that

$$\theta_* = S\beta - s \implies \beta = S^{-1}(\theta_* + s).$$

Then, the criterion function of PMLE will be

$$\begin{aligned} Q(\theta) &= \frac{1}{2} \log \sigma^2 + \frac{1}{2n\sigma^2} (Y - Z\beta)' (Y - Z\beta) \\ &= \frac{1}{2} \log \sigma^2 + \frac{1}{2n\sigma^2} (Y - ZS^{-1}(\theta_* + s))' (Y - ZS^{-1}(\theta_* + s)). \end{aligned}$$

The first order conditions are

$$\begin{aligned} Q_{\theta_*}(\tilde{\theta}) &= -\frac{1}{n\tilde{\sigma}^2} (S^{-1}Z') (Y - ZS^{-1}(\tilde{\theta}_* + s)) \\ Q_{\sigma^2}(\tilde{\theta}) &= \frac{1}{\tilde{\sigma}^2} - \frac{1}{n\tilde{\sigma}^4} (Y - ZS^{-1}(\tilde{\theta}_* + s))' (Y - ZS^{-1}(\tilde{\theta}_* + s)). \end{aligned}$$

So, we have that

$$\tilde{\sigma}^2 = \frac{1}{n} (Y - Z\tilde{\beta})' (Y - Z\tilde{\beta}).$$

Also

$$Q_{\theta_*\theta_*}(\bar{\theta}) = \frac{1}{\tilde{\sigma}^2} \left( S^{-1} \frac{Z'Z}{n} S'^{-1} \right) \xrightarrow{P} \mathcal{D}^{-1}.$$

Therefore,

$$\begin{aligned}\mathcal{LM} &= nQ_{\theta_*'}(\tilde{\theta})Q_{\theta_*\theta_*}^{-1}(\bar{\theta})Q_{\theta_*}(\tilde{\theta}) \\ &= (Y - Z\tilde{\beta})' \frac{Z(Z'Z)^{-1}Z'}{\tilde{\sigma}^2} (Y - Z\tilde{\beta}) \\ &= nR^2,\end{aligned}$$

because  $Y - Z\tilde{\beta}$  are the residuals least squares. That is, the LM is  $n$  times the coefficient of multiple correlation in the regression of  $\tilde{u}_i$  on  $z_i$ . That is, if the least squares estimate in that regression model is denoted by  $\tilde{\alpha}$ , then

$$R^2 = \frac{\tilde{\alpha}'Z'Z\tilde{\alpha}}{\tilde{u}'\tilde{u}}.$$

**Example 30** *Note that*

$$\begin{aligned} Q(\hat{\theta}) &= \frac{1}{2} \log \hat{\sigma}^2 + \frac{1}{2n} \frac{(Y - Z\hat{\beta})'(Y - Z\hat{\beta})}{\hat{\sigma}^2} \\ &= \frac{1}{2} \log \hat{\sigma}^2 + \frac{1}{2}, \end{aligned}$$

*because*

$$\hat{\sigma}^2 = \frac{1}{n} (Y - Z\hat{\beta})'(Y - Z\hat{\beta}).$$

*Similarly,*

$$\begin{aligned} Q(\tilde{\theta}) &= \frac{1}{2} \log \tilde{\sigma}^2 + \frac{1}{2n} \frac{(Y - Z\tilde{\beta})'(Y - Z\tilde{\beta})}{\tilde{\sigma}^2} \\ &= \frac{1}{2} \log \tilde{\sigma}^2 + \frac{1}{2}, \end{aligned}$$

where  $\tilde{\beta}$  is the restricted least squares estimator of  $\beta$ .

*Hence, we conclude that in this case*

$$\mathcal{LR} = n \log \left( \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right).$$

Now,

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{1}{n} (Y - Z\tilde{\beta})' (Y - Z\tilde{\beta}) \\ &= \hat{\sigma}^2 + (\hat{\beta} - \tilde{\beta})' \widehat{M} (\hat{\beta} - \tilde{\beta})\end{aligned}$$

because  $Z$  is orthogonal to  $\widehat{U}$ .

On the other hand, the restricted least squares estimator  $\tilde{\beta}$  is

$$\tilde{\beta} = \hat{\beta} + \widehat{M}^{-1}R' (R\widehat{M}^{-1}R')^{-1} (r - R\hat{\beta}).$$

Then, we obtain that

$$\begin{aligned}\tilde{\sigma}^2 &= \hat{\sigma}^2 + (R\hat{\beta} - r)' (R\widehat{M}^{-1}R')^{-1} (R\hat{\beta} - r) \\ &= \hat{\sigma}^2 + \frac{\mathcal{W}}{n} \hat{\sigma}^2,\end{aligned}$$

That is,

$$\mathcal{W} = \frac{n(\tilde{\sigma}^2 - \hat{\sigma}^2)}{\hat{\sigma}^2},$$

while

$$\mathcal{LR} = n \log \left( \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right) = \frac{n (\tilde{\sigma}^2 - \hat{\sigma}^2)}{\bar{\sigma}^2},$$

where  $\bar{\sigma}^2$  is a mean value. Moreover,

$$\mathcal{LM} = \frac{n (\tilde{\sigma}^2 - \hat{\sigma}^2)}{\tilde{\sigma}^2}$$

## 8 Inference under More General Conditions

- For instance, the criterion function  $Q$  may not be differentiable or the true parameter value may not be interior of the parameter space.
- Suppose

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} Q(\theta)$$

and

$$Q(\theta) \xrightarrow{p} \bar{Q}(\theta)$$

uniformly in  $\Theta$ .

- In this case, we need to take a 3-step approach.

1. consistency,  $\hat{\theta} \xrightarrow{p} \theta_0$
2. convergence rate  $r_n$ , i.e.,  $r_n (\hat{\theta} - \theta_0) = O_p(1)$ .
3. weak convergence of reparametrized criterion function

$$r_n^2 \left( Q \left( \theta_0 + \frac{h}{r_n} \right) - Q(\theta_0) \right)$$

or, more generally,

$$\begin{aligned} & r_n^2 \left( Q \left( \theta_0 + \frac{h}{r_n} \right) - Q(\theta_0) - \left( \bar{Q} \left( \theta_0 + \frac{h}{r_n} \right) - \bar{Q}(\theta_0) \right) \right) \\ & + r_n^2 \left( \bar{Q} \left( \theta_0 + \frac{h}{r_n} \right) - \bar{Q}(\theta_0) \right) \end{aligned}$$

and application of the argmax CMT.

**Theorem 48 (Argmax CMT)** *Let  $X_n, X$  be stochastic processes indexed by a metric space  $\mathcal{T}$  such that  $X_n \Rightarrow X$  in  $\ell^\infty(K)$  for every compact  $K \subset \mathcal{T}$ . Suppose that almost all sample paths of  $X$  are upper semicontinuous and possesses a unique maximum at a (random) point  $\hat{\tau}$ , which is  $O_p(1)$ . If the sequence  $\hat{\tau}_n = O_p(1)$ , then  $\hat{\tau}_n \Rightarrow \hat{\tau}$  in  $\mathcal{T}$ .*

**Definition 26** *A sequence of stochastic processes  $\{\nu_n(\cdot) : n \geq 1\}$  is stochastically (or asymptotically) equicontinuous if  $\forall \varepsilon > 0$  and  $\eta > 0$ ,  $\exists \delta > 0$  and  $n_0$  such that*

$$\sup_{n \geq n_0} \Pr \left\{ \sup_{\rho(\tau_1, \tau_2) < \delta} \|\nu_n(\tau_1) - \nu_n(\tau_2)\| > \eta \right\} < \varepsilon,$$

*for all large  $n$ .*

## 9 ASYMPTOTICS WHEN THE TRUE VALUE OF THE PARAMETER IS AT THE BOUNDARY

- The purpose of this section is to present the basic ideas and results of extreme estimators when the true value,  $\theta_0$  say, may lie at the boundary of the compact parameter space  $\Theta \subset \mathbb{R}^p$ . We shall begin with the simplest of the models, see for instance Gouriéroux, Holly and Monfort (1982) *Econometrica* paper. However, we shall proceed quite differently than them.
- Suppose that we have the following linear regression model

$$y_i = \beta x_i + u_i; \quad i = 1, \dots, n.$$

It is assumed that  $\Theta \equiv [0, M]$ , for some  $M < \infty$ . If  $\{u_i\}_{i \in \mathbb{Z}}$  were a sequence of *iid* random variables, the obvious objective function would be

$$Q(\beta) = \sum_{i=1}^n (y_i - \beta x_i)^2,$$

although even if  $\{u_i\}_{i \in \mathbb{Z}}$  were heteroscedastic and/or autocorrelated we may still employ  $Q(\beta)$  as objective function to estimate  $\beta$ .

- Obviously, our estimator of  $\beta$  is given by

$$\widehat{\beta} = \arg \min_{\beta \in \Theta} Q(\beta), \quad (8.1)$$

and we know that under suitable regularity conditions,  $\widehat{\beta} \rightarrow_P \beta_0$ . In particular, we need that

$$\Pr \left\{ \sup_{\beta \notin \mathcal{N}(\beta_0)} Q(\beta_0) - Q(\beta) > 0 \right\} \rightarrow 0, \quad (8.2)$$

where  $\mathcal{N}(\beta_0) = \{\beta : |\beta - \beta_0| \leq \varepsilon\}$ . That (8.2) holds true comes from the

(obvious) observation that

$$\begin{aligned}
 Q(\beta_0) - Q(\beta) &= -(\beta - \beta_0)^2 \frac{1}{n} \sum_{i=1}^n x_i^2 \\
 &\quad + 2(\beta_0 - \beta) \frac{1}{n} \sum_{i=1}^n x_i u_i \\
 &\xrightarrow{p} -(\beta - \beta_0)^2 E x_i^2
 \end{aligned} \tag{8.3}$$

uniformly in  $\beta$  and that  $-(\beta - \beta_0)^2 E x_i^2 < 0$ .

- Next, we shall see the rate of convergence, that is the value of  $\alpha$  for which (8.2) still true but where now

$$\mathcal{N}(\beta_0) = \{\beta : |\beta - \beta_0| \leq L n^{-\alpha}\}$$

for some  $L > 0$ . To that end and denoting  $\beta_0 - \beta$  by  $\tilde{\beta}$ , we need to show that, using the first equality in (8.3),

$$\Pr \left\{ \sup_{\mathcal{A}_n(\varepsilon)} -\tilde{\beta}^2 \frac{X'X}{n} + \frac{2\tilde{\beta}}{n} X'U > 0 \right\} < \eta \tag{8.4}$$

for some arbitrarily small  $\eta$ , and where

$$\mathcal{A}_n(\varepsilon) = \left\{ \varepsilon > |\beta - \beta_0| \geq Ln^{-1/2} \right\}.$$

- The left side of (8.4) is

$$\begin{aligned} & \Pr \left\{ \sup_{\mathcal{A}_n(\varepsilon)} -\tilde{\beta}^2 + 2\tilde{\beta} \left( \frac{X'X}{n} \right)^{-1} \frac{X'U}{n} > 0 \right\} \\ & \leq \Pr \left\{ \sup_{\mathcal{A}_n(\varepsilon)} -|\tilde{\beta}| + 2 \frac{\tilde{\beta}}{|\tilde{\beta}|} \left( \frac{X'X}{n} \right)^{-1} \frac{X'U}{n} > 0 \right\} \\ & \leq \Pr \left\{ 2 \left( \frac{X'X}{n} \right)^{-1} \left| \frac{X'U}{n} \right| > \inf_{\mathcal{A}_n(\varepsilon)} |\tilde{\beta}| \right\} \\ & \leq \Pr \left\{ 2 \left( \frac{X'X}{n} \right)^{-1} \left| \frac{X'U}{n^{1/2}} \right| > L \right\} \end{aligned}$$

which is bounded by  $\eta$  choosing  $L$  large enough, because  $n^{-1/2} \sum_{i=1}^n x_i u_i = O_p(1)$  and  $n^{-1} \sum_{i=1}^n x_i^2$  converges in probability to  $Ex_i^2 > 0$ .

- Next, the previous result indicates that our estimator of  $\beta_0$  given in (8.1) can be written as

$$\widehat{\beta} = \beta_0 + \frac{\widehat{v}}{n^{1/2}} \quad (8.5)$$

where

$$\widehat{v} = \arg \min_{|v| < L} Q\left(\beta_0 + \frac{v}{n^{1/2}}\right).$$

Also, the previous arguments indicate that, when minimizing  $Q(\beta)$ , we only need to consider  $\beta$ 's of the type

$$\beta = \beta_0 + \frac{v}{n^{1/2}},$$

with  $|v| < L \in (0, \infty)$ .

- Let us consider

$$n(Q(\beta_0) - Q(\beta)) = n\left(Q(\beta_0) - Q\left(\beta_0 + \frac{v}{n^{1/2}}\right)\right).$$

Note that the problem has become the maximization problem. From (8.3),

we have that the right side of the last displayed equation is

$$-v^2 \frac{1}{n} \sum_{i=1}^n x_i^2 - 2v \frac{1}{n^{1/2}} \sum_{i=1}^n x_i u_i := \Lambda_n(v).$$

So, we can consider  $\Lambda_n(v)$  as a process indexed by  $v$ , where  $v$  belongs to a compact set. Moreover, the process is continuous in  $v$ , that is  $\Lambda_n(v) \in \mathbb{C}[-L, L]$ . We shall now see where  $\Lambda_n(v)$  converges.

- To that end, we need to check two things. (a) The convergence of the finite-dimensional distributions and (b) that the process is tight. We shall begin with (a). It is clear that for any finite collection  $v_1, \dots, v_q$ ,

$$\begin{pmatrix} \Lambda_n(v_1) \\ \cdot \\ \cdot \\ \cdot \\ \Lambda_n(v_q) \end{pmatrix} \xrightarrow{d} - \begin{pmatrix} v_1^2 \\ \cdot \\ \cdot \\ \cdot \\ v_q^2 \end{pmatrix} \sigma_x^2 + 2\Omega^{1/2} \mathcal{N}(0, \sigma_u^2 \sigma_x^2)$$

where  $\sigma_x^2 = E(x_i^2)$  and the  $(i, j)$ th element of  $\Omega$  is  $v_i v_j$ . In particular if  $q = 1$ , we have that

$$\Lambda_n(v) \xrightarrow{d} -v^2 \sigma_x^2 + 2v \sigma_u \sigma_x \mathcal{Z}$$

where  $\mathcal{Z}$  denotes the standard normal random variable.

- Next, we need to show tightness, that is (b). Consider  $\bar{\Lambda}_n(v_2, v_1) = \Lambda_n(v_2) - \Lambda_n(v_1)$  for any  $v_2 > v_1$ . By Theorem 12.6 of Billingsley (1968), a sufficient condition for tightness is that

$$E \left| \bar{\Lambda}_n(v_2, v_1) - E \bar{\Lambda}_n(v_2, v_1) \right|^\xi < K (F(v_2) - F(v_1))^{1+\delta} \quad (8.6)$$

for some  $\xi > 0$ ,  $\delta > 0$  and where  $F(\cdot)$  is a monotonic nondecreasing and

continuous function. By definition of  $\Lambda_T(v)$ ,

$$\begin{aligned}
 & E \left| \bar{\Lambda}_n(v_2, v_1) - E \bar{\Lambda}_n(v_2, v_1) \right|^\xi \\
 = & E \left| v_2 \frac{1}{n^{1/2}} \sum_{i=1}^n x_i u_i - v_1 \frac{1}{n^{1/2}} \sum_{i=1}^n x_i u_i \right|^\xi \\
 \leq & (v_2 - v_1)^\xi E \left| \frac{1}{n^{1/2}} \sum_{i=1}^n x_i u_i \right|^\xi \\
 \leq & D (v_2 - v_1)^\xi,
 \end{aligned}$$

which implies that (8.6) holds true choosing  $\xi > 2$ . It goes without saying that I am assuming that the  $\xi$ th moment of  $x_i u_i$  is finite.

- So, we have shown that

$$\Lambda_n(v) \xrightarrow{\text{weakly}} \Lambda(v) \equiv -v^2 \sigma_x^2 + 2v \sigma_u \sigma_x \mathcal{Z}.$$

Now, the limit process  $\Lambda(v)$  is a parabola with fixed second derivatives in  $v$ . So, if this is the case, it turns out that the functional “arg max” is

continuous, so that by the continuous mapping theorem we can conclude that

$$\widehat{v} = \arg \max_{|v| < L} \Lambda_n(v) \xrightarrow{d} v^* = \arg \max_{|v| < L} \Lambda(v). \quad (8.7)$$

But, by definition of  $\Lambda_n(v)$ , we have that

$$n^{1/2} \left( \widehat{\beta} - \beta_0 \right) = \widehat{v}.$$

- Now, what about  $v^*$ ? Suppose that  $\beta_0 > 0$ . Then the true value is an interior point of the parameter space  $\Theta$ . In this case, we have that all the  $\beta$  of the type  $\beta = \beta_0 + n^{-1/2}v$  belong to  $\Theta$  for all  $v$ , so that there is no constraints in  $v$  and thence

$$v^* = \sigma_u \sigma_x^{-1} \mathcal{Z}$$

is the maximum of  $\Lambda(v)$ , and we conclude that

$$n^{1/2} \left( \widehat{\beta} - \beta_0 \right) \xrightarrow{d} \sigma_u \sigma_x^{-1} \mathcal{Z}.$$

This is the case when  $\beta_0 > 0$ .

- Now what about if  $\beta_0 = 0$ ? It is obvious that we cannot expect the same type of asymptotic distribution because  $\widehat{\beta} - \beta_0 \geq 0$ , so that the only "admissible" values of  $v$  are  $\geq 0$ . So,

$$v^* = \sigma_u \sigma_x^{-1} \mathcal{Z} \mathbb{I}(\mathcal{Z} \geq 0)$$

which implies that

$$n^{1/2} \left( \widehat{\beta} - \beta_0 \right) \xrightarrow{d} \sigma_u \sigma_x^{-1} \mathcal{Z} \mathbb{I}(\mathcal{Z} \geq 0)$$

by (8.7).

## 9.1 NONLINEAR MODELS

In this section we discuss what happens with nonlinear models such a nonlinear regression models. Consider the nonlinear regression model

$$y_i = f(x_i; \beta) + u_i; \quad i = 1, \dots, n.$$

Again we assume that  $\Theta \equiv [0, M]$ , for some  $M < \infty$ . If  $\{u_i\}_{i \in \mathbb{Z}}$  were a sequence of *iid* random variables, the obvious objective function would be

$$Q(\beta) = \sum_{i=1}^n (y_i - f_i(\beta))^2,$$

where we have abbreviated  $f(x_i; \beta)$  by  $f_i(\beta)$ . Obviously, our estimator of  $\beta$  is given by

$$\hat{\beta} = \arg \min_{\beta \in \Theta} Q(\beta). \quad (8.8)$$

We know that under suitable regularity conditions,  $\hat{\beta} \xrightarrow{P} \beta_0$ . In particular we

need that

$$\Pr \left\{ \sup_{\beta \notin \mathcal{N}(\beta_0)} Q(\beta_0) - Q(\beta) > 0 \right\} \rightarrow 0, \quad (8.9)$$

where  $\mathcal{N}(\beta_0) = \left\{ \beta : \left| \tilde{\beta} \right| \leq \varepsilon \right\}$ . That (8.9) holds true comes from the observation that

$$\begin{aligned} Q(\beta_0) - Q(\beta) &= -\tilde{\beta}^2 \frac{1}{n} \sum_{i=1}^n (f'_i(\bar{\beta}))^2 \\ &\quad + 2\tilde{\beta} \frac{1}{n} \sum_{i=1}^n f'_i(\bar{\beta}) u_i \\ &\xrightarrow{P} -\tilde{\beta}^2 E(f'_i(\bar{\beta}))^2 \end{aligned} \quad (8.10)$$

uniformly in  $\beta$  since as usual we can give several conditions under which

$$\frac{1}{n} \sum_{i=1}^n (f'_i(\beta))^2 \xrightarrow{P} E(f'_i(\beta))^2$$

uniformly in  $\beta$ .

Next, we shall see the rate of convergence, that is the value of  $\alpha$  for which (8.9) still true but where now  $\mathcal{N}(\beta_0) = \left\{ \beta : \left| \tilde{\beta} \right| \leq Ln^{-\alpha} \right\}$  for some  $L > 0$ . To that end, we need to show that, using the first equality in the last displayed expression,

$$\Pr \left\{ \sup_{\mathcal{A}_n(\varepsilon)} - \frac{\tilde{\beta}^2}{n} F'(\bar{\beta}) F(\bar{\beta}) + 2 \frac{\tilde{\beta}}{n} F'(\bar{\beta}) U > 0 \right\} < \eta \quad (8.11)$$

for some arbitrarily small  $\eta$ , where  $F(\bar{\beta})$  is the matrix out of  $f'_i(\bar{\beta})$ . The proof of (8.11) proceeds as that of (8.4) after we note that

$$\sup_{\beta \in \Theta} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n f'_i(\beta) u_i \right| = O_p(1). \quad (8.12)$$

So, we shall show (8.12). The way we are going to do it is as follows.

Denoting

$$\Upsilon_n(\beta) = \frac{1}{n^{1/2}} \sum_{i=1}^n f'_i(\beta) u_i,$$

we are going to show that

$$\Upsilon_n(\beta) \stackrel{weakly}{\Rightarrow} \Upsilon(\beta),$$

actually to a Gaussian process. As we did in the previous section, we need to show (a) the convergence of the finite dimensional distributions and (b) tightness condition. The proof of (a) is trivial. This is not more than to show a *CLT* property of  $\Upsilon_n(\beta_j)$  for  $j = 1, \dots, q$  with  $q$  finite. Next (b). As we proceed above we shall show that

$$E|\Upsilon_n(\beta_2) - E\Upsilon_n(\beta_1)|^\xi < K(H(\beta_2) - H(\beta_1))^{1+\delta} \quad (8.13)$$

for some  $\xi > 0$ ,  $\delta > 0$  and where  $H(\cdot)$  is a monotonic nondecreasing and continuous function. Take  $\xi = 4$ . Then the left side of (8.13) is

$$\begin{aligned} & \frac{1}{n^2} \sum_{i=1}^n E(f'_i(\beta_2) - f'_i(\beta_1))^4 Eu_i^4 \\ & + \frac{6}{n^2} \sum_{i < j} E \left\{ (f'_i(\beta_2) - f'_i(\beta_1))^2 (f'_j(\beta_2) - f'_j(\beta_1))^2 \right\} \\ & \quad \times Eu_i^2 Eu_j^2. \end{aligned}$$

Now assuming that  $|f'_i(\beta_2) - f'_i(\beta_1)| \leq |\beta_2 - \beta_1|^\varsigma g_i(\beta)$  with  $E \sup_{\beta \in \Theta} |g_i(\beta)|^4 < K$ , then it is easily shown that the last displayed expression is bounded by

$$\frac{K}{n} |\beta_2 - \beta_1|^{4\varsigma} + K |\beta_2 - \beta_1|^{4\varsigma}$$

which implies that if  $\varsigma > \frac{1}{4}$ , then (8.13) holds true with  $\delta = 4\varsigma - 1$ . So, we have that

$$\Upsilon_n(\beta) \xrightarrow{\text{weakly}} \Upsilon(\beta)$$

and since the “sup” is a continuous function we have that (8.12) holds true. From here the proof proceeds as in the case of the linear regression model but replacing  $x_i$  by  $f'_i(\beta)$ . The only difference is to notice that

$$\frac{1}{n^{1/2}} \sum_{i=1}^n f'_i \left( \beta_0 + \tau \frac{v}{n^{1/2}} \right) u_i$$

where  $\tau \in (0, 1)$ , satisfies that its difference with

$$\frac{1}{n^{1/2}} \sum_{i=1}^n f'_i(\beta_0) u_i,$$

that is,

$$\frac{1}{n^{1/2}} \sum_{i=1}^n \left( f'_i \left( \beta_0 + \tau \frac{v}{n^{1/2}} \right) - f'_i(\beta_0) \right) u_i$$

satisfies

$$\sup_{|v| < L} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \left( f'_i \left( \beta_0 + \tau \frac{v}{n^{1/2}} \right) - f'_i(\beta_0) \right) u_i \right| = o_p(1).$$

The latter displayed equality implies that instead of

$$\frac{1}{n^{1/2}} \sum_{i=1}^n f'_i \left( \beta_0 + \tau \frac{v}{n^{1/2}} \right) u_i$$

we only need to consider  $n^{-1/2} \sum_{i=1}^n f'_i(\beta_0) u_i$ .

## References

- [1] BLANCHARD, O. J. and QUAH, D. (1989), “The Dynamic Effects of Aggregate Demand and Supply Disturbances”, *American Economic Review*, 79, 654-673.
- [2] Christiano, L., and Eichenbaum, M. and Evans, C. L. (1999), Monetary policy shocks: What have we learned and to what end?, ch. 02, p. 65-148 in Taylor, J. B. and Woodford, M. eds., *Handbook of Macroeconomics*, vol. 1, Part A, Elsevier.
- [3] GALÍ, J. (1992), “How Well Does the IS-LM Model Fit Postwar U.S. Data?” *Quarterly Journal of Economics*, 107 (2), 709–738.
- [4] Manski, C. F. (1988): Identification of Binary Response Models, *Journal of the American Statistical Association*, 83: 403, pp. 729-738.
- [5] ROTHENBERG, T. J. (1971), “Identification in Parametric Models”, *Econometrica*, 39 (3), 577–591.

- [6] RUBIO-RAMÍREZ, J.F., and DANIEL F. WAGGONER, and TAO ZHA (2010). "Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference," *Review of Economic Studies*, 77(2), pages 665-696.
- [7] SIMS, C. A. (1980), "Macroeconomics and Reality", *Econometrica*, 48 (1), 1-48.
- [8] Sims, C.A. (2002), "Structural VARs", Lecture note. ??