METHOD OF MOMENTS

Herman J. Bierens

Pennsylvania State University

September 19, 2005

1. Linear method of moments

1.1. The model

Consider a system of k linear equations,

$$y_{i,t} = x_{i,t}^T\theta_i + u_{i,t},\ t = 1,...,n,\quad i = 1,...,k,\ \theta_i \in \mathbb{R}^{p_i}, \tag{1}$$

where the $x_{i.,t}$ vectors possibly contain some of the dependent variables $y_{j,t}$, and the errors $u_{i,t}$ have zero expectation but are contemporaneously dependent. Consequently, the usual regression assumption $E[u_{i,t}|x_{i,t}] = 0$ may not apply. However, suppose we have q_i-vectors $z_{i.,t}$ of instrumental variables such that $E[u_{i,t}z_{i,t}] = 0$, hence

$$\left(E[z_{i,t}x_{i,t}^T]\right)\theta_i = E[z_{i,t}y_{i,t}] \tag{2}$$

is a system of q_i linear equations in p_i unknown elements of θ_i. If $q_i \geq p_i$, and the rank of the matrix $E[z_{i,t}x_{i,t}^T]$ is p_i or larger, the parameter vector θ_i is identified by the moment conditions (2).

Note that this case is only one of the many cases for which the method of moment estimation approach is applicable. For example, least squares and two-stage least squares estimation are special cases of method of moment estimation techniques.

Denoting

$$p = \sum_{i=1}^{k} p_i,\quad q = \sum_{i=1}^{k} q_i, \tag{3}$$

we can write this model in vector form as

$$y_t = X_t^T\theta_0 + u_t,\quad E[Z_t u_t] = 0, \tag{4}$$

where

$$y_t = \begin{pmatrix} y_{1,t} \\ \vdots \\ y_{k,t} \end{pmatrix}, \quad X_t = \begin{pmatrix} x_{1,t} & 0 & \dots & 0 \\ 0 & x_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{k,t} \end{pmatrix} \quad (p \times k), \quad \theta_0 = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}, \quad u_t = \begin{pmatrix} u_{1,t} \\ \vdots \\ u_{k,t} \end{pmatrix}, \tag{5}$$

and

$$Z_t = \begin{pmatrix} z_{1,t} & 0 & \dots & 0 \\ 0 & z_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{k,t} \end{pmatrix} \quad (q \times k). \tag{6}$$

The moment conditions (2) now take the form

$$E[Z_t X_t^T]\theta_0 = E[Z_t y_t]. \tag{7}$$

The implicit assumption that the parameter vectors θ_i in model (1) are different is not essential. If some or all of the components of the parameter vectors θ_i are common, we may augment the $x_{i.,t}$ vectors with zeros corresponding to the parameters that are not part of the equation involved, and write model (1) as

$$y_{i,t} = x_{i,t}^T\theta_0 + u_{i,t}, \; t = 1,...,n, \quad i = 1,...,k, \; \theta_0 \in \mathbb{R}^p. \tag{8}$$

The only difference is then that the matrix X_t in (5) becomes

$$X_t = (x_{1,t}, ..., x_{k,t}) \quad (p \times k). \tag{9}$$

The moment conditions (7) suggest to estimate θ_0 by minimizing the quadratic form

$$Q_n(\theta) = M_n(\theta)^T W_n M_n(\theta), \tag{10}$$

where

$$M_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} Z_t\left(y_t - X_t^T\theta\right) = \frac{1}{n}\sum_{t=1}^{n} Z_t y_t - \left(\frac{1}{n}\sum_{t=1}^{n} Z_t X_t^T\right)\theta. \tag{11}$$

and W_n is a positive definite $q \times q$ matrix, to be determined later. Thus, the *method of moment* estimator involved is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, Q_n(\theta). \tag{12}$$

If $q = p$, the solution is the same as the solution of $M_n(\theta) = 0$, namely

$$\hat{\theta} = \left(\frac{1}{n}\sum_{t=1}^{n} Z_t X_t^T\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t y_t\right), \tag{13}$$

provided that the inverted matrix is nonsingular, which is known as the just identified case, but the overidentified case $q > p$ is more interesting and challenging.

The first-order condition for a minimum of $Q_n(\theta)$ is:

$$\frac{\partial Q_n(\theta)}{\partial \theta^T} = 2\left(\frac{\partial M_n(\theta)^T}{\partial \theta^T}\right) W_n M_n(\theta) = -2\left(\frac{1}{n}\sum_{t=1}^{n} X_t Z_t^T\right) W_n\left(\frac{1}{n}\sum_{t=1}^{n} Z_t y_t - \frac{1}{n}\sum_{t=1}^{n} Z_t X_t^T\theta\right) = 0 \tag{14}$$

hence the solution is:

$$\begin{aligned}\hat{\theta} &= \left[\left(\frac{1}{n}\sum_{t=1}^{n} X_t Z_t^T\right) W_n\left(\frac{1}{n}\sum_{t=1}^{n} Z_t X_t^T\right)\right]^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} X_t Z_t^T\right) W_n\left(\frac{1}{n}\sum_{t=1}^{n} Z_t y_t\right) \\ &= \theta_0 + \left[\left(\frac{1}{n}\sum_{t=1}^{n} X_t Z_t^T\right) W_n\left(\frac{1}{n}\sum_{t=1}^{n} Z_t X_t^T\right)\right]^{-1}\cdot\left(\frac{1}{n}\sum_{t=1}^{n} X_t Z_t^T\right) W_n\left(\frac{1}{n}\sum_{t=1}^{n} Z_t u_t\right)\end{aligned} \tag{15}$$

Now assume that

Assumption 1: $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t u_t \to N_k(0, A)$ *in distr., where* $A = \operatorname{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} Z_t u_t u_t^T Z_t^T$.

This condition is satisfied if for example $Z_t u_t$ is i.i.d. and the variance matrix A of $Z_t u_t$ is finite. Moreover, assume that

Assumption 2: $B = \text{plim}_{n\to\infty}(1/n)\sum_{t=1}^{n} X_t Z_t^T$ *exists and is finite,*

Assumption 3: $W = \text{plim}_{n\to\infty} W_n$ *is finite and positive definite*.

Assumption 4: BWB^T *is nonsingular*.

Then under Assumptions 1-4,

$$\sqrt{n}(\hat{\theta} - \theta) \to N_p[0\ ,\ (BWB^T)^{-1}(BWAWB^T)(BWB^T)^{-1}]\ \textit{in distr.} \tag{16}$$

Because the variance matrix involved depends on *W*, the question now arises:

1.2. What is the best choice for W_n ?

In order to answer this question, consider the linear regression model

$$y = A^{-1/2}B^T\theta + e,\ e \sim N(0,I). \tag{17}$$

It follows from the Gauss-Markov theorem that the best linear unbiased estimator of θ is the least squares estimator:

$$\hat{\theta} = (BA^{-1}B^T)^{-1}BA^{-1/2}y = \theta + (BA^{-1}B^T)^{-1}BA^{-1/2}e \sim N[\theta,(BA^{-1}B^T)^{-1}] \tag{18}$$

Next, consider the alternative unbiased estimator

$$\begin{aligned}\tilde{\theta} &= [(BWB^T)^{-1}BWA^{1/2}]y = \theta + [(BWB^T)^{-1}BWA^{1/2}]e \\ &\sim N[\theta,(BWB^T)^{-1}(BWAWB^T)(BWB^T)^{-1}]\end{aligned} \tag{19}$$

By the Gauss-Markov theorem,

$$D = (BWB^T)^{-1}(BWAWB^T)(BWB^T)^{-1} - (BA^{-1}B^T)^{-1} \tag{20}$$

is a positive semi-definite matrix. The direct proof follows from the fact that we can write

$$D = \left[(BWB^T)^{-1}BWA^{1/2}\right]\left(I - A^{-1/2}B^T\left(BA^{-1}B^T\right)^{-1}BA^{-1/2}\right)\left[A^{1/2}WB^T(BWB^T)^{-1}\right] \tag{21}$$

and that the matrix $I - A^{-1/2}B^T\left(BA^{-1}B^T\right)^{-1}BA^{-1/2}$ is idempotent, hence positive semi-definite.

Since $D = O$ if $W = A^{-1}$, the best choice for W_n is therefore such that

$$\operatorname{plim}_{n\to\infty} W_n = A^{-1}. \tag{22}$$

The *efficient method of moment estimation* procedure is now as follows. First, choose an initial matrix W_n, for example, let $W_n = I_q$. Then compute the first stage method of moment estimator $\hat{\theta}$, and denote

$$\hat{A} = \frac{1}{n}\sum_{t=1}^{n} Z_t\hat{u}_t\hat{u}_t^T Z_t^T, \quad \textit{where } \hat{u}_t = y_t - X_t^T\hat{\theta}. \tag{23}$$

Next, choose

$$W_n = \hat{A}^{-1}, \tag{24}$$

which under Assumptions 1-4 is a consistent estimator of A^{-1}. Using this matrix W_n in the second stage now yields the *efficient method of moment* estimator:

$$\hat{\theta}_{EMM} = \operatorname*{argmin}_{\theta} M_n(\theta)^T\hat{A}^{-1}M_n(\theta), \tag{25}$$

with limiting normal distribution:

$$\sqrt{n}\left(\hat{\theta}_{EMM} - \theta_0\right) \to N_p[0, (BA^{-1}B^T)^{-1}]. \tag{26}$$

Moreover, denoting

$$\hat{B} = \frac{1}{n}\sum_{t=1}^{n} X_t Z_t^T, \tag{27}$$

it follows from Assumptions 1,2, and 4 that the asymptotic variance matrix in (26) can be estimated consistently by $(\hat{B}\hat{A}^{-1}\hat{B}^T)^{-1}$.

1.3. Testing the adequacy of the instruments

It follows from (15) with (24) and (27) that

$$\sqrt{n}(\hat{\theta}_{EMM} - \theta_0) = \left(\hat{B}\hat{A}^{-1}\hat{B}^T\right)^{-1}\left(\hat{B}\hat{A}^{-1}\right)\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t u_t\right), \tag{28}$$

hence it follows from (11) and Assumption 1 that

$$\begin{aligned}\sqrt{n}\hat{A}^{-1/2}M_n(\hat{\theta}_{EMM}) &= \hat{A}^{-1/2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t u_t - \hat{A}^{-1/2}\hat{B}^T\sqrt{n}(\hat{\theta}_{EMM} - \theta_0) \\ &= \left[\hat{A}^{-1/2} - \hat{A}^{-1/2}\hat{B}^T\left(\hat{B}\hat{A}^{-1}\hat{B}^T\right)^{-1}\left(\hat{B}\hat{A}^{-1}\right)\right]\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t u_t\right) \\ &= \left[I_q - \hat{A}^{-1/2}\hat{B}^T\left(\hat{B}\hat{A}^{-1}\hat{B}^T\right)^{-1}\left(\hat{B}\hat{A}^{-1/2}\right)\right]\left(\hat{A}^{-1/2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t u_t\right) \\ &\rightarrow N_q[0, M] \text{ in distribution},\end{aligned} \tag{29}$$

where

$$M = I_q - A^{-1/2}B^T(BA^{-1}B^T)^{-1}BA^{-1/2}. \tag{30}$$

Since M is idempotent (*Exercise*: Why?), it follows now that under Assumptions 1-4,

$$nM_n(\hat{\theta}_{EMM})^T\hat{A}^{-1}M_n(\hat{\theta}_{EMM}) \rightarrow \chi^2_{q-p} \text{ in distribution}. \tag{31}$$

(*Exercise*: Why?) On the other hand, if

$$\operatorname{plim}_{n\rightarrow\infty}\frac{1}{n}\sum_{t=1}^{n} Z_t u_t = \eta \neq 0, \tag{32}$$

and $q > p$ then under Assumptions 2-4,

$$\operatorname{plim}_{n\rightarrow\infty}\hat{A}^{-1/2}M_n(\hat{\theta}_{EMM}) = \left[A^{-1/2} - A^{-1/2}B^T\left(BA^{-1}B^T\right)^{-1}\left(BA^{-1}\right)\right]\eta \neq 0 \tag{33}$$

hence

$$\operatorname{plim}_{n\rightarrow\infty} n.M_n(\hat{\theta}_{EMM})^T\hat{A}^{-1}M_n(\hat{\theta}_{EMM}) = \infty. \tag{34}$$

Therefore, we can use $nM_n(\hat{\theta}_{EMM})^T\hat{A}^{-1}M_n(\hat{\theta}_{EMM})$ as a test for the adequacy of the instruments.

1.4. Application to static panel data models

A static panel data model takes the form[1]

$$y_{i,t}^{*} = x_{i,t}^{*T}\theta_0 + \alpha_i + \varepsilon_{i,t},\ t = 1,...,T,\quad i = 1,...,N,\ \theta_0 \in \mathbb{R}^p, \tag{35}$$

where α_i is a fixed or random effect which is constant over time t but varies with the cross-section index i, the $x_{i,t}^{*}$ are $p\times 1$ vectors of exogenous variables, **none** of which are constant over time, and the $\varepsilon_{i,t}$ are i.i.d. $(0,\sigma^2)$ errors that are independent of the exogenous variables. It will be assumed that the cross-section dimension N is much larger than the time dimension T, so that T may be considered as fixed, whereas all the asymptotic properties follow from letting $N \to \infty$.

In order to get rid of α_i, we take first differences:

$$\begin{aligned} y_{i,t}^{*} - y_{i,t-1}^{*} &= (x_{i,t}^{*} - x_{i,t-1}^{*})^T\theta_0 + \varepsilon_{i,t} - \varepsilon_{i,t-1}, \\ t &= 2,...,T,\quad i = 1,...,N,\ \theta_0 \in \mathbb{R}^p, \end{aligned} \tag{36}$$

Since $\varepsilon_{i,t} - \varepsilon_{i,t-1}$ is independent of $x_{i,t}^{*}$ and $x_{i,t-1}^{*}$, we can choose either $x_{i,t}^{*} - x_{i,t-1}^{*}$ or $x_{i,t}^{*}$ and $x_{i,t-1}^{*}$ as instruments. Choosing the latter, we can now write the model in vector form as

$$y_i = X_i^T\theta_0 + u_i,\quad E[Z_i u_i] = 0,\ i = 1,....,N, \tag{37}$$

where

$$y_i = \begin{pmatrix} y_{i,2}^{*} - y_{i,1}^{*} \\ \vdots \\ y_{i,T}^{*} - y_{i,T-1}^{*} \end{pmatrix},\quad X_i = (x_{i,2}^{*} - x_{i,1}^{*},...,x_{i,T}^{*} - x_{i,T-1}^{*}),\quad u_i = \begin{pmatrix} \varepsilon_{i,2}^{*} - \varepsilon_{i,1}^{*} \\ \vdots \\ \varepsilon_{i,T}^{*} - \varepsilon_{i,T-1}^{*} \end{pmatrix}, \tag{38}$$

and

[1] Note that T now denotes the length of the time series, whereas the superscript ^{T} still denotes the "transpose".

$$Z_t = \begin{pmatrix} x_{i,1}^* & 0 & \dots & 0 \\ x_{i,2}^* & 0 & \dots & 0 \\ 0 & x_{i,2}^* & \dots & 0 \\ 0 & x_{i,3}^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{i,T-1}^* \\ 0 & 0 & \dots & x_{i,T}^* \end{pmatrix} \quad (q \times k),\ q = 2p,\ k = T-1. \tag{39}$$

Note that

$$Var(u_i) = \sigma^2 \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} = \sigma^2 \Omega,\ say. \tag{40}$$

Hence

$$A = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} Z_i u_i u_i^T Z_t^T = \sigma^2 \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} Z_i \Omega Z_t^T. \tag{41}$$

Therefore, if we choose

$$W_N = \left(\frac{1}{N} \sum_{i=1}^{N} Z_i \Omega Z_t^T \right)^{-1} \tag{42}$$

as the weight matrix, we obtain the efficient method of moment estimator $\hat{\theta}_{EMM}$ in one step. The only difference with the general case is that $N.M_N(\hat{\theta}_{EMM})^T W_N M_N(\hat{\theta}_{EMM})$ needs to be divided by a consistent estimate $\hat{\sigma}^2$ of the variance σ^2 of the errors $\varepsilon_{i,t}$'s in order to be used as a chi-square

test of model correctness. Of course, we also need $\hat{\sigma}^2$ to estimate the asymptotic variance matrix of $\sqrt{N}(\hat{\theta}_{EMM} - \theta_0)$. For example, given the residual vector $\hat{u}_i = y_i - X_i^T\hat{\theta}_{EMM}$, the variance σ^2 can be estimated consistently by

$$\hat{\sigma}^2 = \frac{1}{2N(T-1)}\sum_{i=1}^{N} \hat{u}_i^T\hat{u}_i. \tag{43}$$

(*Exercise*: Why?)

1.5. Application to dynamic panel data models

A dynamic panel data model takes the form

$$y_{i,t}^* = \rho_0 y_{i,t-1}^* + x_{i,t}^{*T}\beta_0 + \alpha_i + \varepsilon_{i,t},\ t = 2,...,T,\quad i = 1,...,N,\ \beta_0 \in \mathbb{R}^p,\ |\rho_0| < 1, \tag{44}$$

where again α_i is a fixed or random effect which is constant over time t but varies with the cross-section index i, the $x_{i,t}^*$ are $p\times 1$ vectors of exogenous variables which are not constant over time, and the $\varepsilon_{i,t}$ are i.i.d. $(0,\sigma^2)$ errors that are independent of the exogenous variables. Also now it will be assumed that the cross-section dimension N is much larger than the time dimension T, so that T may be considered as fixed, whereas all the asymptotic properties follow from letting $N \to \infty$.

Taking first differences yields for $t = 3,...,T,\quad i = 1,...,N$,

$$y_{i,t}^* - y_{i,t-1}^* = \rho_0(y_{i,t-1}^* - y_{i,t-2}^*) + (x_{i,t}^* - x_{i,t-1}^*)^T\beta_0 + \varepsilon_{i,t} - \varepsilon_{i,t-1}. \tag{45}$$

Due to the dynamic structure of the model, we now have a much richer choice of instruments, because $\rho_0(y_{i,t-1}^* - y_{i,t-2}^*) + (x_{i,t}^* - x_{i,t-1}^*)^T\beta_0$ depends on $y_{i,t-2-j}^*$ for $j = 0,...,t-2$ as well as on $x_{i,t-j}^*$ for $j = 0,...,t$. Denoting

$$X_i = (y_{i,2}^* - y_{i,1}^*, x_{i,3}^* - x_{i,2}^*, ..., y_{i,T-1}^* - y_{i,T-2}^*, x_{i,T}^* - x_{i,T-1}^*) \tag{46}$$

$$\theta_0 = \begin{pmatrix}\rho_0\\ \beta_0\end{pmatrix},\quad y_i = \begin{pmatrix} y_{i,3}^* - y_{i,2}^* \\ \vdots \\ y_{i,T}^* - y_{i,T-1}^*\end{pmatrix},\quad u_i = \begin{pmatrix} \varepsilon_{i,3}^* - \varepsilon_{i,2}^* \\ \vdots \\ \varepsilon_{i,T}^* - \varepsilon_{i,T-1}^*\end{pmatrix}, \tag{47}$$

$$Z_t = \begin{pmatrix} x_{i,1}^* & x_{i,1}^* & \dots & x_{i,1}^* \\ x_{i,2}^* & x_{i,2}^* & \dots & x_{i,2}^* \\ x_{i,3}^* & x_{i,3}^* & \dots & x_{i,3}^* \\ y_{i,1}^* & x_{i,4}^* & \dots & x_{i,4}^* \\ 0 & y_{i,1}^* & \dots & x_{i,5}^* \\ 0 & y_{i,2}^* & \dots & x_{i,6}^* \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{i,T}^* \\ 0 & 0 & \dots & y_{i,1}^* \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & y_{i,T-2}^* \end{pmatrix} \quad (q \times k),\ q = pT+T-1,\ k = T-2, \tag{48}$$

we can write the model again as (37). Therefore, the same results as in the previous section hold, except that we have to modify (43) to

$$\hat{\sigma}^2 = \frac{1}{2N(T-2)} \sum_{i=1}^{N} \hat{u}_i^T \hat{u}_i. \tag{49}$$

(*Exercise*: Why?)

2. Nonlinear method of moments

2.1 The model

Consider now the case where a model for a random vector $X_t \in \mathbb{R}^k$ is implicitly defined by a set of moment conditions:

$$m_t(\theta) = \begin{pmatrix} \mu_1(X_t, \theta) \\ \vdots \\ \mu_q(X_t, \theta) \end{pmatrix}, \quad \theta \in \Theta \subset \mathbb{R}^p, \quad \exists \theta_0 \in \Theta\colon E[m_t(\theta_0)] = 0, \tag{50}$$

where $q \geq p$, Θ is the parameter space, θ_0 is the parameter vector of interest. The random vectors

X_t are observable for $t = 1,..,n$. For convenience of the exposition we will assume that

Assumption A: *the X_t's are i.i.d.,*

but under some mild additional conditions the results below will also hold if the X_t's are realizations of a stationary vector time series process, or are panel data observations.

The following assumptions allow us to apply the central limit theorem and the uniform law of large numbers:

Assumption B: *The functions $\mu_i(x,\theta)$ are twice continuously differentiable in θ, and for each $\theta \in \Theta$ Borel measurable in $x \in \mathbb{R}^k$. The parameter space Θ is compact and convex, and θ_0 is an interior point of Θ.*

Assumption C: *For $i = 1,...,k$,*

$$E\left(\sup_{\theta\in\Theta} \mu_i(X_t,\theta)^2\right) < \infty, \quad E\left(\sup_{\theta\in\Theta} \left\|\frac{\partial \mu_i(X_t,\theta)}{\partial \theta^T}\right\|\right) < \infty, \quad E\left(\sup_{\theta\in\Theta} \left\|\frac{\partial^2 \mu_i(X_t,\theta)}{\partial \theta \partial \theta^T}\right\|\right) < \infty.$$

In the latter case one should interpret the matrix norm $\|\cdot\|$ as the maximum of the absolute values of the elements of the matrix involved. Moreover, in order for the parameter vector θ_0 to be identified, we need to ensure that

Assumption D: $\|E[m_t(\theta)]\| = 0$ *if and only if* $\theta = \theta_0$.

2.2. Strong consistency

Denote

$$M_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} m_t(\theta), \quad \bar{M}(\theta) = E[m_t(\theta)]. \tag{51}$$

Under Assumptions A-C we have

$$\sup_{\theta\in\Theta} \|M_n(\theta) - \bar{M}(\theta)\| \to 0 \ a.s., \tag{52}$$

hence, denoting

$$Q_n(\theta) = M_n(\theta)^T W_n M_n(\theta), \quad \bar{Q}(\theta) = \bar{M}(\theta)^T W \bar{M}(\theta), \tag{53}$$

where

$$W_n \to W \ a.s., \quad with \ W \ a \ positive \ definite \ symmetric \ matrix, \tag{54}$$

it follows that

$$\sup_{\theta\in\Theta} |Q_n(\theta) - \bar{Q}(\theta)| \to 0 \ a.s.. \tag{55}$$

This result, together with Assumption D, imply that

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, Q_n(\theta) \to \theta_0 \ a.s. \tag{56}$$

(*Exercise*: Why?)

3.3. Asymptotic normality

Assumptions A-C are also sufficient conditions for the application of the central limit theorem:

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} m_t(\theta_0) \to N_q[0, A] \ in \ distr., \ where \ A = E[m_t(\theta_0) m_t(\theta_0)^T]. \tag{57}$$

The limiting normal distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ can be derived as follows. The first-order conditions for a minimum of $Q_n(\theta)$ are:

$$\frac{\partial Q_n(\theta)}{\partial \theta_i} = 2\left(\frac{\partial M_n(\theta)^T}{\partial \theta_i}\right) W_n M_n(\theta) = 2\left(\frac{1}{n}\sum_{t=1}^{n} \frac{\partial m_t(\theta)^T}{\partial \theta_i}\right) W_n M_n(\theta) = 0 \tag{58}$$

for $i = 1,...,p$. If $\hat{\theta}$ is on the border of the parameter space Θ, these first-order conditions may not

hold for $\hat{\theta}$. However, since by Assumption B θ_0 is an interior point of Θ, and $\hat{\theta} \to \theta_0$ a.s., we have that

$$P\left[\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n M_n(\theta)\Big|_{\theta=\hat{\theta}} = 0\right] \to 1 \tag{59}$$

Next observe that by the mean value theorem, and the convexity of the parameter space Θ, there exists a mean value $\tilde{\theta}_i \in \Theta$, with $\|\tilde{\theta}_i - \theta_0\| \leq \|\hat{\theta} - \theta_0\|$, such that for $i = 1,...,p$,

$$\begin{aligned}\sqrt{n}\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n M_n(\theta)\Big|_{\theta=\hat{\theta}} &= \sqrt{n}\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n M_n(\theta)\Big|_{\theta=\theta_0} \\ &+ \frac{\partial}{\partial\theta}\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n M_n(\theta)\Big|_{\theta=\tilde{\theta}_i}\sqrt{n}\left(\hat{\theta} - \theta_0\right).\end{aligned} \tag{60}$$

It follows from (59) that the left-hand side of (60) converges in probability to zero. Moreover, since

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n M_n(\theta) &= \left((1/n)\Sigma_{t=1}^{n}\frac{\partial^2 m_t(\theta)^T}{\partial\theta_i\partial\theta_j}\right)W_n M_n(\theta) \\ &+ \left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_i)\right)W_n\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)/\partial\theta_j)\right) \\ \to \left(E\left[\frac{\partial^2 m_t(\theta)^T}{\partial\theta_i\partial\theta_j}\right]\right)W\bar{M}(\theta) &+ \left(E(\partial m_t(\theta)^T/\partial\theta_i)\right)W\left(E(\partial m_t(\theta)/\partial\theta_j)\right) \text{ a.s., uniformly on } \Theta,\end{aligned} \tag{61}$$

and $\tilde{\theta}_i \to \theta_0$ a.s., it follow that

$$\tilde{C} = \begin{pmatrix} \frac{\partial}{\partial\theta}\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_1)\right)W_n M_n(\theta)\Big|_{\theta=\tilde{\theta}_1} \\ \vdots \\ \frac{\partial}{\partial\theta}\left((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta_p)\right)W_n M_n(\theta)\Big|_{\theta=\tilde{\theta}_p} \end{pmatrix} \to BWB^T \text{ a.s.}, \tag{62}$$

(*Exercise*: Why?) where

$$B = E\left[\left.\frac{\partial m_t(\theta)}{\partial \theta}\right|_{\theta=\theta_0}\right]. \tag{63}$$

Furthermore, it follows from the strong law of large numbers that

$$(1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta)\Big|_{\theta=\theta_0} \to B \ a.s. \tag{64}$$

hence it follows from (54) and (57) that

$$\sqrt{n}\Big((1/n)\Sigma_{t=1}^{n}(\partial m_t(\theta)^T/\partial\theta)\Big)W_n M_n(\theta)\Big|_{\theta=\theta_0} \to N_p(0, BWAWB^T) \ \textit{in distr.} \tag{65}$$

(*Exercise*: Why?) Combining the results (59), (60), (62), and (65), now yield (*Exercise*: Why?)

$$\sqrt{n}(\hat{\theta} - \theta) \to N_p[0\ ,\ (BWB^T)^{-1}(BWAWB^T)(BWB^T)^{-1}] \ \textit{in distr.} \tag{66}$$

provided that BWB^T is nonsingular. Again, the variance matrix involved is the smallest for $W = A^{-1}$, provided of course that

Assumption E: *The matrix $BA^{-1}B^T$ is nonsingular.*

Thus, $W_n = \hat{A}^{-1}$ is an optimal choice.

Finally, observe that under Assumptions A-D,

$$\hat{A} = \frac{1}{n}\sum_{t=1}^{n} m_t(\theta)m_t(\theta)^T \Big|_{\theta=\hat{\theta}} \to A \ a.s., \quad \hat{B} = \frac{1}{n}\sum_{t=1}^{n}\left.\frac{\partial m_t(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}} \to B \ a.s.. \tag{67}$$

(*Exercise*: Why?) Thus, under Assumptions A-E the *efficient method of moment* estimator

$$\hat{\theta}_{EMM} = \underset{\theta}{\operatorname{argmin}}\, M_n(\theta)^T\hat{A}^{-1}M_n(\theta), \tag{68}$$

is strongly consistent: $\hat{\theta}_{EMM} \to \theta_0$ a.s., and has limiting normal distribution:

$$\sqrt{n}\left(\hat{\theta}_{EMM} - \theta_0\right) \to N_p[0, (BA^{-1}B^T)^{-1}]. \tag{69}$$

Moreover,

$$(\hat{B}\hat{A}^{-1}\hat{B}^{T})^{-1} \rightarrow (BA^{-1}B^{T})^{-1} \ a.s. \tag{70}$$

Finally, observe that similarly to the linear case, under Assumptions A-E and the null hypothesis H_0: $E[m_t(\theta_0)] = 0$ for some $\theta_0 \in \Theta$,

$$nM_n(\hat{\theta}_{EMM})^T\hat{A}^{-1}M_n(\hat{\theta}_{EMM}) \rightarrow \chi^2_{q-p} \ in \ distr., \tag{71}$$

whereas under the alternative hypothesis that the null hypothesis is incorrect, and the maintained Assumptions A, B,C,E,

$$nM_n(\hat{\theta}_{EMM})^T\hat{A}^{-1}M_n(\hat{\theta}_{EMM}) \rightarrow \infty \ a.s. \tag{72}$$

Thus, the left-hand side of (71) is the test statistic of the Wald test that the moment conditions involved are correct.