STOR 754: Summary Course Notes: Supplement 2

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
rls@email.unc.edu

October 3, 2008

Course Text: *Time Series: Theory and Methods* (Springer Series in Statistics) by Peter J. Brockwell and Richard A. Davis. Second edition, 1991.

1 Asymptotics of MLE for ARMA Time Series

First, let us recall the results of Brockwell and Davis (B&D):

Assume X_t is a causal invertible ARMA(p,q) process with parameters $\boldsymbol{\beta}^T = (\ \beta_1 \quad \beta_2 \quad \ldots \quad \beta_{p+q}\ ) = (\ \phi_1 \quad \ldots \quad \phi_p \quad \theta_1 \quad \ldots \quad \theta_q\ )$. We assume the noise sequence $\{\epsilon_t\}$ consists of IID random variables with common mean 0 and variance σ^2. Note that σ^2 is not explicitly included as an unknown parameter though it is often estimated independently once $\boldsymbol{\beta}$ has been determined.

Suppose we have observations $X_1, ..., X_n$ and calculate the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_n$. The main result as stated by B&D is that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N[0, V(\boldsymbol{\beta})]$. To define $V(\boldsymbol{\beta})$, B&D define two auxiliary processes $\{U_t\}$ and $\{V_t\}$ by the equations $\phi(B)U_t = \epsilon_t$, $\theta(B)V_t = \epsilon_t$ (where as usual, $\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$, $\theta(B) = 1 + \theta_1 B + ... + \theta_q B^q$), and let $\mathbf{U} = (\ U_{t-1} \quad \ldots \quad U_{t-p}\ ), \mathbf{V} = (\ V_{t-1} \quad \ldots \quad V_{t-q}\ )$ for some fixed t. Then

$$V(\boldsymbol{\beta}) \quad = \quad \sigma^2 \begin{bmatrix} \mathrm{E}\{\mathbf{U}\mathbf{U}^T\} & \mathrm{E}\{\mathbf{U}\mathbf{V}^T\} \\ \mathrm{E}\{\mathbf{V}\mathbf{U}^T\} & \mathrm{E}\{\mathbf{V}\mathbf{V}^T\} \end{bmatrix}^{-1}. \tag{1}$$

In particular, for an AR(p) process, U_t coincides with X_t and (1) allows the simpler representation

$$V(\boldsymbol{\beta}) \quad = \quad \sigma^2 \Gamma_p^{-1}, \tag{2}$$

where Γ_p is the covariance matrix of $(\ X_1 \quad \ldots \quad X_p\ )$.

B&D also state the result in the following alternative form (p. 386). Suppose the spectral density of X_t is of the form $f(\lambda) = \frac{\sigma^2}{2\pi} g(\lambda)$ where

$$g(\lambda) \quad = \quad \frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2}. \tag{3}$$

The $V(\boldsymbol{\beta}) = W(\boldsymbol{\beta})^{-1}$ where W has entries $\{w_{jk}\}$,

$$w_{jk} \quad = \quad \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \frac{\partial \log g(\lambda; \boldsymbol{\beta})}{\partial \beta_j} \right\} \left\{ \frac{\partial \log g(\lambda; \boldsymbol{\beta})}{\partial \beta_k} \right\} d\lambda. \tag{4}$$

Proof that (1) and (4) are equivalent

Case 1: Suppose $1 \leq j, k \leq p$. So we have

$$\begin{aligned} \frac{\partial \log g(\lambda;\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left\{\log(\phi(e^{i\lambda})) + \log(\phi(e^{-i\lambda})) - \log(\theta(e^{i\lambda})) - \log(\theta(e^{-i\lambda}))\right\} \\ &= \frac{e^{ij\lambda}}{\phi(e^{i\lambda})} + \frac{e^{-ij\lambda}}{\phi(e^{-i\lambda})} \end{aligned}$$

and similary for $\frac{\partial \log g(\lambda;\boldsymbol{\beta})}{\partial \beta_k}$, so the right hand side of (4) becomes

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\frac{e^{ij\lambda}}{\phi(e^{i\lambda})} + \frac{e^{-ij\lambda}}{\phi(e^{-i\lambda})}\right\}\left\{\frac{e^{ik\lambda}}{\phi(e^{i\lambda})} + \frac{e^{-ik\lambda}}{\phi(e^{-i\lambda})}\right\} d\lambda. \tag{5}$$

However, we also have (from the fact that the spectral density of U_t is $\frac{\sigma^2}{2\pi}|\phi(e^{i\lambda})|^{-2}$) that

$$\frac{1}{\sigma^2}\mathrm{E}(U_j U_k) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{i(j-k)\lambda}}{|\phi(e^{i\lambda})|^2} d\lambda. \tag{6}$$

However, (5) and (6) are equivalent because of the result (see B&D, three lines before the end of p. 386)

$$\int_{-\pi}^{\pi} e^{i(j+k)\lambda}\phi^{-2}(e^{i\lambda})d\lambda = \int_{-\pi}^{\pi} e^{-i(j+k)\lambda}\phi^{-2}(e^{-i\lambda})d\lambda = 0. \quad (j, k \geq 1) \tag{7}$$

To see that (7) is true, there are two possible proofs: (a) write $\phi^{-1}(e^{i\lambda}) = \sum_{r=0}^{\infty}\psi_r e^{i\lambda r}$ with $\sum_r |\psi_r| < \infty$; then $\int_{-\pi}^{\pi} e^{i\lambda(j+k)}\sum_r\sum_s \psi_r\psi_s e^{i(r+s)\lambda}d\lambda = 0$ because each $j+k+r+s > 0$; (b) substitute either $e^{i\lambda} = z$ or $e^{-i\lambda} = z$ and rewrite (7) as a contour integral around the unit circle in the complex plane; the integrand becomes $z^{j+k-1}\phi(z)^{-2}$ which is analytic inside the unit circle; hence the integral is 0. Note also that the same proof works if $\phi^{-2}(e^{-i\lambda})$ is replaced by either $\theta^{-2}(e^{-i\lambda})$ or $\phi^{-1}(e^{-i\lambda})\theta^{-1}(e^{-i\lambda})$ (see (10) below).

Case 2: Suppose $j, k > p$, say $j = p + j',\ k = p + k'$ so that (β_j, β_k) correspond to $(\theta_{j'}, \theta_{k'})$. Then, similarly to (5), w_{jk} becomes

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\frac{e^{ij'\lambda}}{\theta(e^{i\lambda})} + \frac{e^{-ij'\lambda}}{\theta(e^{-i\lambda})}\right\}\left\{\frac{e^{ik'\lambda}}{\theta(e^{i\lambda})} + \frac{e^{-ik'\lambda}}{\theta(e^{-i\lambda})}\right\} d\lambda$$

while

$$\frac{1}{\sigma^2}\mathrm{E}(V_{j'}V_{k'}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{i(j'-k')\lambda}}{|\theta(e^{i\lambda})|^2} d\lambda,$$

so the result follows by the same argument as Case 1.

Case 3: Suppose $j \leq p,\ k = p + k' > p$. In this case, w_{jk} is

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\frac{e^{ij\lambda}}{\phi(e^{i\lambda})} + \frac{e^{-ij\lambda}}{\phi(e^{-i\lambda})}\right\}\left\{\frac{e^{ik'\lambda}}{\theta(e^{i\lambda})} + \frac{e^{-ik'\lambda}}{\theta(e^{-i\lambda})}\right\} d\lambda. \tag{8}$$

We need the following small extension of the result at the beginning of Chapter 3 of the notes: suppose X_t has spectral density $f_X(\lambda)$ and $Y_t = \sum c_r X_{t-r},\ Z_t = \sum d_s X_{t-s}$, with $\sum c_r^2 < \infty$, $\sum d_s^2 < \infty$. Also let $C(z) = \sum c_r z^r$, $D(z) = \sum d_s z^s$, convergent on $|z| \leq 1$. Then

$$\mathrm{Cov}(Y_t, Z_{t+k}) = \int_{-\pi}^{\pi} e^{ik\lambda}C(e^{i\lambda})D(e^{-i\lambda})f_X(\lambda)d\lambda \tag{9}$$

From this result we deduce

$$\frac{1}{\sigma^2}\mathrm{E}(U_{t-j}V_{t-k'}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{i(j-k')\lambda}}{\phi(e^{i\lambda})\theta(e^{-i\lambda})}d\lambda. \tag{10}$$

The equality of (8) and (10) follows from the extension of (7) noted earlier.

Before going further, we note the following elementary result: if $\{x_n\}$ are a sequence of real numbers and $x_n \to x$ as $n \to \infty$, then

$$\frac{1}{n}\sum_{t=1}^{n} x_t \to x. \tag{11}$$

Now we turn to our own derivation of the asymptotic result for $\hat{\boldsymbol{\beta}}$.

Assume that the optimal linear predictor of X_t given X_{t-k}, $k = 1, ..., n$ is of the form $\sum_{k=1}^{n}\pi_{n,k}X_{t-k}$ and that the resulting optimal mean squared error is $P_{n,1}$. In particular, for $n = t-1$, we have the optimal predictor $\hat{X}_t = \sum_{k-1}^{t-1}\pi_{t-1,k}X_{t-k}$. Also define the sequence $\{\pi_k\}$ by the expansion $\pi(z) = \sum_{k=0}^{\infty}\pi_k z^k$, $\pi(z) = \frac{\phi(z)}{\theta(z)}$. We claim (proof given later) that as $n \to \infty$,

$$\pi_{n,k} \to \pi_k \text{ for each } k \geq 0, \tag{12}$$

$$P_{n,1} \to \sigma^2. \tag{13}$$

From the prediction error decomposition, the likelihood L_n based on observations $X_1, ..., X_n$ is given by

$$L_n = \prod_{t=1}^{n}\left[(2\pi P_{t-1,1})^{-1/2}\exp\left\{-\frac{1}{2}\frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}}\right\}\right].$$

Ignoring 2π constants, $\ell_n = -\log L_n$ and its derivatives are given by

$$\begin{aligned}
\ell_n &= \sum_{t=1}^{n}\Big[\frac{1}{2}\log P_{t-1,1} + \frac{1}{2}\frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}}\Big],\\
\frac{\partial \ell_n}{\partial \beta_j} &= \sum_{t=1}^{n}\Big[\frac{1}{2}\frac{\partial}{\partial\beta_j}\log P_{t-1,1} + \frac{\partial\hat{X}_t}{\partial\beta_j}\cdot\frac{\hat{X}_t - X_t}{P_{t-1,1}} - \frac{1}{2}\frac{\partial P_{t-1,1}}{\partial\beta_j}\cdot\frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}^2}\Big]\\
\frac{\partial^2\ell_n}{\partial\beta_j\partial\beta_k} &= \sum_{t=1}^{n}\Big[\frac{1}{2}\frac{\partial^2}{\partial\beta_j\partial\beta_k}\log P_{t-1,1} + \frac{\partial^2\hat{X}_t}{\partial\beta_j\partial\beta_k}\cdot\frac{\hat{X}_t - X_t}{P_{t-1,1}} + \frac{\partial\hat{X}_t}{\partial\beta_j}\frac{\partial\hat{X}_t}{\partial\beta_k}\frac{1}{P_{t-1,1}}\\
&\quad -\frac{\partial\hat{X}_t}{\partial\beta_j}\frac{\partial P_{t-1,1}}{\partial\beta_k}\cdot\frac{\hat{X}_t - X_t}{P_{t-1,1}^2} - \frac{1}{2}\frac{\partial^2 P_{t-1,1}}{\partial\beta_j\partial\beta_k}\cdot\frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}^2} - \frac{\partial P_{t-1,1}}{\partial\beta_j}\frac{\partial\hat{X}_t}{\partial\beta_k}\cdot\frac{\hat{X}_t - X_t}{P_{t-1,1}^2}\\
&\quad +\frac{\partial P_{t-1,1}}{\partial\beta_j}\frac{\partial P_{t-1,1}}{\partial\beta_k}\cdot\frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}^3}\Big],\\
\frac{1}{n}\mathrm{E}\left\{\frac{\partial^2\ell_n}{\partial\beta_j\partial\beta_k}\right\} &= \frac{1}{n}\sum_{t=1}^{n}\Big[\frac{1}{2}\frac{\partial^2}{\partial\beta_j\partial\beta_k}\log P_{t-1,1} + \mathrm{E}\left\{\frac{\partial\hat{X}_t}{\partial\beta_j}\frac{\partial\hat{X}_t}{\partial\beta_k}\right\}\frac{1}{P_{t-1,1}}\\
&\quad -\frac{1}{2}\frac{\partial^2 P_{t-1,1}}{\partial\beta_j\partial\beta_k}\cdot\frac{1}{P_{t-1,1}} + \frac{\partial P_{t-1,1}}{\partial\beta_j}\frac{\partial P_{t-1,1}}{\partial\beta_k}\cdot\frac{1}{P_{t-1,1}^2}\Big].
\end{aligned} \tag{14}$$

Since $P_{t-1,1} \to \sigma^2$ (independent of $\boldsymbol{\beta}$) as $t \to \infty$, and since $P_{t-1,1}$ is continuously differentiable with respect to $\boldsymbol{\beta}$, it follows that the partial derivatives of $P_{t-1,1}$ and $\log P_{t-1,1}$ with respect to components of $\boldsymbol{\beta}$ tend to 0 as $t \to \infty$, so by (11), the first, third and fourth terms in (14) all tend to 0. For the second term, we have

$$\begin{aligned} \mathrm{E}\left\{\frac{\partial \hat{X}_t}{\partial \beta_j}\frac{\partial \hat{X}_t}{\partial \beta_k}\right\} &= \sum_{r=1}^{t-1}\sum_{s=1}^{t-1}\frac{\partial \pi_{t,r}}{\partial \beta_j}\frac{\partial \pi_{t,s}}{\partial \beta_k}\gamma_X(r-s) \\ &\to \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{\partial \pi_r}{\partial \beta_j}\frac{\partial \pi_s}{\partial \beta_k}\gamma_X(r-s). \end{aligned}$$

Hence we find (again using (11)) that

$$\frac{1}{n}\mathrm{E}\left\{\frac{\partial^2 \ell_n}{\partial \beta_j \partial \beta_k}\right\} \to \frac{1}{\sigma^2}\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{\partial \pi_r}{\partial \beta_j}\frac{\partial \pi_s}{\partial \beta_k}\gamma_X(r-s). \tag{15}$$

First, we note that if X_t is AR(p), then (15) leads directly to the Brockwell-Davis formula. For in that case, $\pi_j = \phi_j = \beta_j$ for $j = 1, ..., p$ (and $\pi_j = 0$ for $j > p$) so (15) is exactly $\frac{1}{\sigma^2}\gamma_X(j-k)$. Hence the normalized Fisher information matrix is $\frac{1}{\sigma^2}\Gamma_p$ and its inverse is $\sigma^2\Gamma_p^{-1}$, as in (1).

For the general case (15), we have

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{\partial \pi_r}{\partial \beta_j}\frac{\partial \pi_s}{\partial \beta_k}\gamma_X(r-s) &= \frac{1}{\sigma^2}\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\frac{\partial \pi_r}{\partial \beta_j}\frac{\partial \pi_s}{\partial \beta_k}\int_{-\pi}^{\pi} e^{i(r-s)\lambda} f_X(\lambda)d\lambda \\ &= \frac{1}{\sigma^2}\int_{-\pi}^{\pi}\frac{\partial}{\partial \beta_j}\left\{\sum_{r=1}^{\infty}\pi_r e^{ir\lambda}\right\}\frac{\partial}{\partial \beta_k}\left\{\sum_{s=1}^{\infty}\pi_s e^{-is\lambda}\right\} f_X(\lambda)d\lambda \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\partial}{\partial \beta_j}\left\{\frac{\phi(e^{i\lambda})}{\theta(e^{i\lambda})}\right\}\frac{\partial}{\partial \beta_k}\left\{\frac{\phi(e^{-i\lambda})}{\theta(e^{-i\lambda})}\right\}\frac{\theta(e^{i\lambda})\theta(e^{-i\lambda})}{\phi(e^{i\lambda})\phi(e^{-i\lambda})}d\lambda \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\partial}{\partial \beta_j}\left\{\log\phi(e^{i\lambda}) - \log\theta(e^{i\lambda})\right\}\frac{\partial}{\partial \beta_k}\left\{\log\phi(e^{-i\lambda}) - \log\theta(e^{-i\lambda})\right\}d\lambda \end{aligned}$$

We break this up into three cases, similarly to (4).

If $1 \le j, k \le p$ then

$$\begin{aligned} \frac{1}{n}\mathrm{E}\left\{\frac{\partial^2 \ell_n}{\partial \beta_j \partial \beta_k}\right\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{ij\lambda}}{\phi(e^{i\lambda})}\cdot\frac{e^{-ik\lambda}}{\phi(e^{-i\lambda})}d\lambda \\ &= \frac{1}{\sigma^2}\mathrm{E}U_{t-j}U_{t-k}. \end{aligned} \tag{16}$$

If $j, k > p$, say $j = p + j',\ k = p + k'$, then

$$\begin{aligned} \frac{1}{n}\mathrm{E}\left\{\frac{\partial^2 \ell_n}{\partial \beta_j \partial \beta_k}\right\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{ij'\lambda}}{\theta(e^{i\lambda})}\cdot\frac{e^{-ik'\lambda}}{\theta(e^{-i\lambda})}d\lambda \\ &= \frac{1}{\sigma^2}\mathrm{E}V_{t-j'}V_{t-k'}. \end{aligned} \tag{17}$$

For the third case, if $j \le p,\ k = p + k' > p$, then by (10),

$$\begin{aligned} \frac{1}{n}\mathrm{E}\left\{\frac{\partial^2 \ell_n}{\partial \beta_j \partial \beta_k}\right\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{ij\lambda}}{\phi(e^{i\lambda})}\cdot\frac{e^{-ik'\lambda}}{\theta(e^{-i\lambda})}d\lambda \\ &= \frac{1}{\sigma^2}\mathrm{E}U_{t-j}V_{t-k'}. \end{aligned} \tag{18}$$

This completes the proof that the limiting Fisher information matrix is given by (1).

Proof of (12) and (13)

Given the sequence $n = \{1, 2, 3, ..., \}$, suppose we have a subsequence n' on which $\pi_{n',k}$ converges for every $k \geq 0$, say $\pi_{n',k} \to \pi_k^*$. It will suffice to show $\pi_k^* = \pi_k$ for each k.

First, we note that

$$\mathrm{E}\left(X_t - \sum_{k=1}^{\infty} \pi_k^* X_{t-k}\right)^2 \quad \geq \quad \sigma^2. \tag{19}$$

(19) follows because we earlier showed that the optimal predictor of X_t given the infinite past has mean square prediction error σ^2 (which is achieved when $\pi_k^* = \pi_k$ for all k).

However, we can also try the solution

$$\tilde{\pi}_{n,k} \quad = \quad \begin{cases} \pi_k & \text{if } k \leq n, \\ 0 & \text{if } k > n. \end{cases}$$

From earlier results about L^2 convergence we have that

$$\begin{aligned} \lim_{n\to\infty} \mathrm{E}\left(X_t - \sum_{k=1}^{n} \tilde{\pi}_{n,k} X_{t-k}\right)^2 \quad &= \quad \mathrm{E}\left(X_t - \sum_{k=1}^{\infty} \pi_k X_{t-k}\right)^2 \\ &= \quad \sigma^2. \end{aligned}$$

Thus by Fatou's Lemma,

$$\begin{aligned} \mathrm{E}\left(X_t - \sum_{k=1}^{\infty} \pi_k^* X_{t-k}\right)^2 \quad &= \quad \mathrm{E} \lim_{n\to\infty} \left(X_t - \sum_{k=1}^{n} \pi_{n,k} X_{t-k}\right)^2 \\ &\leq \quad \liminf_{n\to\infty} \mathrm{E}\left(X_t - \sum_{k=1}^{n} \pi_{n,k} X_{t-k}\right)^2 \\ &\leq \quad \liminf_{n\to\infty} \mathrm{E}\left(X_t - \sum_{k=1}^{n} \tilde{\pi}_{n,k} X_{t-k}\right)^2 \\ &= \quad \sigma^2. \end{aligned} \tag{20}$$

However, comparing (19) and (20), it follows that $\mathrm{E}\left(X_t - \sum_{k=1}^{\infty} \pi_k^* X_{t-k}\right)^2 \; = \; \sigma^2$, and hence, from our earlier optimality result about prediction given the infinite past, that $\pi_k^* = \pi_k$ for all k. This completes the proof of (12) and (13).

1.1 Proof of Asymptotic Normality

So far, we have shown that the Fisher information matrix based on observations $X_1, ..., X_n$ is asymptotically of the from $nV(\boldsymbol{\beta})^{-1}$, where $V(\boldsymbol{\beta})$ is given by (1). It remains to be shown that this actually implies asymptotic normality of $\hat{\boldsymbol{\beta}}$ with asymptotic covariance matrix $n^{-1}V(\boldsymbol{\beta})$. For this we use *Sweeting's Theorem*, a very general result on asymptotic normality of the maximum likelihood estimator due to Sweeting (1980).

Following Sweeting, suppose $\ell_n(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathcal{B}$, for some $\mathcal{B} \subseteq \mathcal{R}^k$, is the negative log likelihood of a parameter vector $\boldsymbol{\beta}$ based on observations $X_1, ..., X_n$, and let $\mathcal{I}_n(\boldsymbol{\beta})$ denote the matrix of second-order derivatives of $\ell_n(\boldsymbol{\beta})$ — this is the (random) observed information matrix. Sweeting also assumes there exist non-random square matrices $A_n(\boldsymbol{\beta})$ where $A_n(\boldsymbol{\beta})^{-1} \to 0$ uniformly in compact subsets of $\mathcal{B}$, and such that, for any fixed $c > 0$,

$$W_n(\boldsymbol{\beta}) \quad = \quad A_n(\boldsymbol{\beta}')^{-1}\mathcal{I}_n(\Gamma)[A_n(\boldsymbol{\beta}')^{-1}]^T \xrightarrow{u} W(\boldsymbol{\beta}), \tag{21}$$

where

- $\boldsymbol{\beta}'$ satisfies $|A_n(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta})| \leq c$,
- if Γ is any matrix $(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)$, $\boldsymbol{\beta}_i \in \mathcal{B}$ for $i = 1, ..., k$, then $\mathcal{I}_n(\Gamma)$ is $\mathcal{I}_n$ with the ith row evaluated at $\boldsymbol{\beta}_i$,
- $\Pr\{W(\boldsymbol{\beta}) > 0\} = 1$.

Then, with probability tending to 1 as $n \to \infty$, there exists $\hat{\boldsymbol{\beta}}_n$ locally minimizing $\ell_n(\boldsymbol{\beta})$, such that

$$W_n^{1/2} A_n^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N[0, I_k]. \tag{22}$$

In (21), $\xrightarrow{u}$ refers to *uniform weak convergence*: if $P_{n,\boldsymbol{\beta}}$ is a sequence of probability measures indexed by $\boldsymbol{\beta} \in \mathcal{B}$, then $P_{n,\boldsymbol{\beta}} \xrightarrow{u} P_{\boldsymbol{\beta}}$ if $\int u dP_{n,\boldsymbol{\beta}} \to \int u dP_{\boldsymbol{\beta}}$ for all real bounded uniformly continuous functions u.

In the case of time series, it is possible to set $A_n = n^{1/2} I_k$ (I_k is the $k \times k$ identity matrix, where here $k = p + q$), $W(\boldsymbol{\beta}) = V(\boldsymbol{\beta})^{-1}$, then the result (22) becomes

$$\sqrt{n} V(\boldsymbol{\beta})^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N[0, I_k]. \tag{23}$$

The remaining task is therefore to verify (21) in this case. This means, in effect, establishing a law of large numbers for $\frac{\partial^2 \ell_n}{\partial \beta_j \partial \beta_k}$.

From (14) we have

$$\begin{aligned}
\frac{1}{n}\frac{\partial^2 \ell_n}{\partial \beta_j \partial \beta_k} \quad = \quad & \frac{1}{n}\sum_{t=1}^{n}\Bigg[\frac{1}{2}\frac{\partial^2}{\partial \beta_j \partial \beta_k}\log P_{t-1,1} + \frac{\partial^2 \hat{X}_t}{\partial \beta_j \partial \beta_k} \cdot \frac{\hat{X}_t - X_t}{P_{t-1,1}} + \frac{\partial \hat{X}_t}{\partial \beta_j}\frac{\partial \hat{X}_t}{\partial \beta_k}\frac{1}{P_{t-1,1}} \\
& - \frac{\partial \hat{X}_t}{\partial \beta_j}\frac{\partial P_{t-1,1}}{\partial \beta_k} \cdot \frac{\hat{X}_t - X_t}{P_{t-1,1}^2} - \frac{1}{2}\frac{\partial^2 P_{t-1,1}}{\partial \beta_j \partial \beta_k} \cdot \frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}^2} - \frac{\partial P_{t-1,1}}{\partial \beta_j}\frac{\partial \hat{X}_t}{\partial \beta_k} \cdot \frac{\hat{X}_t - X_t}{P_{t-1,1}^2} \\
& + \frac{\partial P_{t-1,1}}{\partial \beta_j}\frac{\partial P_{t-1,1}}{\partial \beta_k} \cdot \frac{(\hat{X}_t - X_t)^2}{P_{t-1,1}^3}\Bigg].
\end{aligned} \tag{24}$$

Since we have already calculated the mean of (24), to complete the result it will suffice to show that the variance of (24) tends to 0, uniformly in $\boldsymbol{\beta}$.

We shall only outline how to do this, and since the main contribution to the mean comes from the third of the 7 terms in (24), we shall concentrate on that here, i.e. on showing that the variance of

$$\frac{1}{n}\sum_{t=1}^{n}\frac{1}{P_{t-1,1}}\frac{\partial \hat{X}_t}{\partial \beta_j}\frac{\partial \hat{X}_t}{\partial \beta_k} \tag{25}$$

tends to 0 as $n \to \infty$. In principle, the other terms in (24) follow from similar arguments.

So, let's concentrate on (25). First, we note the following:

Lemma. If $Y_1, ..., Y_4$ are jointly normally distributed with covariances of the form $\mathrm{Cov}(Y_i, Y_j) = \sigma_{ij}$, $1 \leq i, j \leq 4$, then $\mathrm{Cov}(Y_i Y_j, Y_k Y_\ell) = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}$.

Proof. WLOG the means are all 0. If $\mathbf{Y}$ has positive definite covariance matrix Σ, then we can write $\Sigma = AA^T$ for some matrix A (for example, A could be the Cholesky decomposition of Σ, in which case, A is lower triangular, though the triangularity property is not needed for the following discussion). In this case we can write $\mathbf{Y} = A\mathbf{Z}$, where $\mathbf{Z} = (\ Z_1 \quad Z_2 \quad Z_3 \quad Z_4 \ )^T$, $Z_1, ..., Z_4$ being independent $N[0,1]$. Note that $\mathrm{E}(\mathbf{Y}\mathbf{Y}^T) = \mathrm{E}(A\mathbf{Z}\mathbf{Z}^T A^T) = AA^T = \Sigma$.

However with this notation (and writing a_{ij} for the (i,j) entry of A), we have

$$\begin{aligned}
\mathrm{E}(Y_i Y_j Y_k Y_\ell) &= \mathrm{E}\left(\sum_r \sum_s \sum_t \sum_u a_{ir} a_{js} a_{kt} a_{\ell u} Z_r Z_s Z_t Z_u \right) \\
&= 3 \sum_r a_{ir} a_{jr} a_{kr} a_{\ell r} + \sum_{r,s:\ r \neq s} (a_{ir} a_{jr} a_{kr} a_{\ell s} + a_{ir} a_{js} a_{kr} a_{\ell s} + a_{ir} a_{js} a_{ks} a_{\ell r}) \\
&= \sum_r \sum_s (a_{ir} a_{jr} a_{kr} a_{\ell s} + a_{ir} a_{js} a_{kr} a_{\ell s} + a_{ir} a_{js} a_{ks} a_{\ell r}) \\
&= \sigma_{ij}\sigma_{k\ell} + \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}. \qquad (26)
\end{aligned}$$

In the middle of (26), we used the fact that

$$\mathrm{E}\left(Z_r Z_s Z_t Z_u\right) = \begin{cases} 3 & \text{if } r = s = t = u, \\ 1 & \text{if } r = s \neq t = u \text{ or } r = t \neq s = u \text{ or } r = u \neq s = t, \\ 0 & \text{otherwise.} \end{cases}$$

The result of the Lemma then follows from (26) and the fact that $\mathrm{E}(Y_i Y_j)\mathrm{E}(Y_k Y_\ell) = \sigma_{ij}\sigma_{k\ell}$. *QED.*

We now return to the proof that (25) has variance tending to 0 as $n \to \infty$. We can approximate (25) by

$$\frac{1}{n\sigma^2} \sum_{t=1}^{n} \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \frac{\partial \pi_r}{\partial \beta_j} X_{t-r} \frac{\partial \pi_s}{\partial \beta_k} X_{t-s}.$$

The variance of this quantity is

$$\begin{aligned}
& \frac{1}{n^2\sigma^4} \sum_{t=1}^{n} \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \sum_{t'=1}^{n} \sum_{r'=1}^{\infty} \sum_{s'=1}^{\infty} \frac{\partial \pi_r}{\partial \beta_j} \frac{\partial \pi_s}{\partial \beta_k} \frac{\partial \pi_{r'}}{\partial \beta_j} \frac{\partial \pi_{s'}}{\partial \beta_k} \mathrm{Cov}\left(X_{t-r} X_{t-s}, X_{t'-r'} X_{t'-s'}\right) \\
= \ & \frac{1}{n^2\sigma^4} \sum_{t=1}^{n} \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \sum_{t'=1}^{n} \sum_{r'=1}^{\infty} \sum_{s'=1}^{\infty} \frac{\partial \pi_r}{\partial \beta_j} \frac{\partial \pi_s}{\partial \beta_k} \frac{\partial \pi_{r'}}{\partial \beta_j} \frac{\partial \pi_{s'}}{\partial \beta_k} \Big\{ \gamma_X(t' - r' - t + r)\gamma_X(t' - s' - t + s) \\
& \qquad + \gamma_X(t' - r' - t + s)\gamma_X(t' - s' - t + r) \Big\}. \qquad (27)
\end{aligned}$$

We split this up into the same three cases as earlier, according to whether j and k are $\leq p$ or $> p$, but in this case we give the argument only when $1 \leq j, k \leq p$, since the other two cases are similar.

We use results such as

$$\begin{aligned}
\sum_r \sum_{s'} \frac{\partial \pi_r}{\partial \beta_j} \frac{\partial \pi_{s'}}{\partial \beta_k} \gamma_X(t' - s' - t + r) &= \int \frac{\partial \pi(e^{i\lambda})}{\partial \beta_j} \frac{\partial \pi(e^{-i\lambda})}{\partial \beta_k} e^{i\lambda(t'-t)} f_X(\lambda) d\lambda \\
&= \frac{\sigma^2}{2\pi} \int \frac{e^{ij\lambda}}{\phi(e^{i\lambda})} \frac{e^{-ik\lambda}}{\phi(e^{-i\lambda})} e^{i\lambda(t'-t)} d\lambda \\
&= \frac{\sigma^2}{2\pi} \gamma_U(t' - t + j - k)
\end{aligned}$$

to reduce (27) to

$$\frac{1}{4\pi^2 n^2} \sum_{t=1}^{n} \sum_{t'=1}^{n} \Big[\gamma_U(t' - t)^2 + \gamma_U(t' - t + j - k)\gamma_U(t' - t - j + k) \Big]. \qquad (28)$$

It then easily follows that (28) tends to 0 as $n \to \infty$. For instance, the first term in (28) is $\frac{1}{4\pi^2 n^2}\sum_{r=-n+1}^{n-1}(n-|r|)\gamma_U(r)^2 \leq \frac{1}{4\pi^2 n}\sum_{r=\infty}^{\infty}\gamma_U(r)^2 = \frac{K}{n}$ with K a constant. The second term is bounded similarly using the inequality $\gamma_U(t'-t+j-k)\gamma_U(t'-t-j+k) \leq \frac{1}{2}(\gamma_U(t'-t+j-k)^2 + \gamma_U(t'-t-j+k)^2)$. Moreover, the upper bound on the variance is independent of $\boldsymbol{\beta}$, so the convergence is uniform as $\boldsymbol{\beta}$ varies over compact spaces.

This argument establishes that (25) converges to 0 in probability, uniformly over compact subsets of $\boldsymbol{\beta}$. We do not calculate explicitly the remaining terms in (24), but similar arguments are applicable to those terms.

1.2 References

Sweeting, T.J. (1980), Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics*, **8**, 1375–1381.