15.3 Prediction Error Methods

Exercise 3.3 (11.1) On the use of cross-correlation test for the LS model.

Consider an ARX model

$$A(q^{-1})y_t = B(q^{-1})u_t + e_t$$

with parameters $\theta = \left(a_1, \ldots, a_{n_a}, b_1, \ldots b_{n_b}\right)^T$. Assume that the estimate $\hat{\theta}$ is found by LS, show that $\hat{r}_{eu}(\tau) = \frac{1}{n}\sum_{t=1}^{n} u_t e_t = 0$ for all $\tau = 1, \ldots, n_b$.

Solution: The estimate is given by the normal equations

$$\left(\sum_{t=1}^{n} \varphi_t \varphi_t^T\right)\hat{\theta} = \left(\sum_{t=1}^{n} \varphi_t y_t\right)$$

This gives using $\epsilon_t = y_t - \varphi_t^T \hat{\theta}$ that

$$\sum_{t=1}^{n} \varphi_t \epsilon_t = \sum_{t=1}^{n} \varphi_t \left(\varphi_t^T \hat{\theta} - y_t\right) = 0.$$

Exercise 3.4 (11.2) Identifiability results for ARX models.

Consider the system (with A_0, B_0 coprime, and degrees n_{a_0}, n_{b_0})

$$A_0(q^{-1})Y_t = B_0(q^{-1})u_t + D_t$$

where D_t is zero mean, white noise. Let $\{y_t\}_t$ be realizations of the stochastic process $\{Y_t\}$. Use the LS in the model structure

$$A(q^{-1})y_t = B(q^{-1})u_t + \epsilon_t$$

with degrees $n_a \geq n_{a_0}$ and $n_b \geq n_{b_0}$. Assume the system operates in open-loop and the input u_t is PE of order n_b. Let $\epsilon_t(\theta)$ denote the random variables modeling the residuals ϵ_t for given parameters $\theta = (A, B)$. Prove the following results:

(a) The asymptotic cost function $\mathbb{E}[\epsilon_t^2(\theta)]$ has a unique minimum.

(b) The estimated polynomials are coprime.

(c) The information matrix is nonsingular.

Compare with the properties of ARMAX models, see e.g. Example 11.6.

Solution: The prediction error is given as

$$\epsilon_t = A(q^{-1})y_t - B(q^{-1})u_t = \frac{A(q^{-1})B_0(q^{-1}) - A_0(q^{-1})B(q^{-1})}{A_0(q^{-1})}u_t + \frac{A(q^{-1})}{A_0(q^{-1})}D_t$$

Hence the asymptotic loss function satisfies

$$\mathbb{E}[\epsilon_t^2(\theta)] = \mathbb{E}\left[\frac{A(q^{-1})B_0(q^{-1}) - A_0(q^{-1})B(q^{-1})}{A_0(q^{-1})}u_t\right]^2 + \mathbb{E}\left[\frac{A(q^{-1})}{A_0(q^{-1})}D_t\right]^2$$
$$\geq \mathbb{E}\left[\frac{A(q^{-1})}{A_0(q^{-1})}D_t\right]^2 = \mathbb{E}\left[\frac{A(q^{-1})}{A_0(q^{-1})}D_t^2\right] \geq \mathbb{E}\left[D_t^2\right]$$

Hence, equality holds only if

$$\mathbb{E}\frac{AB_0 - A_0B}{A_0}u_t = 0$$

and

$$\frac{A}{A_0} = 1.$$

The second equation gives $A = A_0$, the second one thus $(B_0 - B)u_t = 0$. We have thus proven that the global minimum is unique, and that A, B are coprime.

(c) According to example 11.6, p435 in the book we have that the information matrix is nonsingular if and only if $\varphi_t = \frac{\partial \epsilon_t}{\partial \theta}$ has a nonsingular covariance matrix. Then a relevant equation is

$$\alpha(q^{-1})y_t - \beta(q^{-1})u_t = 0$$

which gives

$$\frac{\alpha(q^{-1})B_0(q^{-1}) - A_0(q^{-1})\beta(q^{-1})}{A_0(q^{-1})}u_t + \frac{\alpha(q^{-1})}{A_0(q^{-1})}D_t = 0.$$

Again, when assuming independence of D_t and u_t, and that for some t $D_t \neq 0$ and u_t is PE of any order we have that $\alpha = \beta = 0$, and hence the information matrix is nonsingular asymptotically (see also eqs. 11.29-11.32) in the book.

Exercise 3.5 (11.7): An illustration of the Parsimony Principle.

Consider the following AR(1) process:

$$Y_t + a_0 Y_{t-1} = D_t,$$

with $|a_0| < 1$ unknown, and where $\{D_t\}_t$ is a white noise stochastic process with zero mean and variance λ^2. Let us have a realization $\{y_t\}_{t=1}^n$ of length n of this process. Then assume we fit this system with the following two candidate models:

$$\mathcal{M}_1 : y_t + a y_{t-1} = \epsilon_t$$

and

$$\mathcal{M}_2 : y_t + a_1 y_{t-1} + a_2 y_{t-2} = \epsilon'_t$$

Let $\hat{a}$ denote the LS estimate of a in $\mathcal{M}_1$, and let $\hat{a}_1, \hat{a}_2$ be the LS estimate in $\mathcal{M}_2$. What are the asymptotic variances $\sqrt{n}(\hat{a} - a_0)$ and $\sqrt{n}(\hat{a}_1 - a_0)$?

Solution: Consider the expression for the inverse of a symmetric 2×2 matrix with $a, c \neq 0$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

The asymptotic covariance matrix of the LS estimate of the parameters of an AR model structure which contains the true system follows the general theory, so we have that

$$P_1 = \lim_{n\to\infty} \frac{n}{\lambda^2} \mathbb{E}[\hat{a} - a_0]^2 = \left(\mathbb{E}[y_t^2]\right)^{-1}$$

For M_2 the covariance is given as

$$P_2 = \lim_{n\to\infty} \frac{n}{\lambda^2} \mathbb{E}[(a - \hat{a})(a - \hat{a})^T] = \left(\begin{bmatrix} \mathbb{E}[Y_t^2] & \mathbb{E}[Y_t Y_{t-1}] \\ \mathbb{E}[Y_t Y_{t-1}] & \mathbb{E}[Y_t^2] \end{bmatrix} \right)^{-1}$$

where we define $r_0 = \mathbb{E}[Y_t^2]$ and $r_1 = \mathbb{E}[Y_t Y_{t-1}]$. This means that $\mathrm{var}(\hat{a}) = \frac{1}{r_0}$ and $\mathrm{var}\, \hat{a}_1 = \frac{r_0}{r_0^2 - r_1^2}$, hence

$$\frac{\mathrm{var}(\hat{a})}{\mathrm{var}(\hat{a}_1)} = 1 - \frac{r_1^2}{r_0^2} \leq 1$$

In order to get more insight in the inequality above, note that some simple calculation shows that for all $k \geq 0$ that

$$r_k = \lambda^2 \frac{(-a_0)^k}{1 - a_o^2}$$

Thus

$$\frac{\mathrm{var}(\hat{a})}{\mathrm{var}(\hat{a}_1)} = 1 - a_0^2,$$

and thus the closer a_0 goes to zero, the smaller the difference between the variances. Note also that for $|a_0|$ close to 1, variance $\mathrm{var}(\hat{a})$ might take very small values, while $\mathrm{var}(\hat{a}_1) = \frac{1}{\lambda^2}$ does not depend on a_0.

Exercise 3.6 (11.8): The parsimony principle does not necessarily hold for nonhierarchical models.

Consider the system

$$\mathcal{S}_1 : \; Y_t = b_0 U_{t-1} + D_t$$

where $\{D_t\}$ and $\{U_t\}_t$ are mutually independent white noise sequences. The variance of D_t is λ^2. Let $\{y_t\}$ and $\{u_t\}$ be realizations of length n of the process $\{Y_t\}, \{U_t\}$ respectively. Consider the following two model structures:

$$\begin{cases} \mathcal{M}_1 : \; y_t + a y_{t-1} = b u_{t-1} + \epsilon_t^1 \\ \mathcal{M}_2 : \; y_t = b_1 u_{t-1} + b_2 u_{t-1} + b_2 u(t-3) + \epsilon_t^2 \end{cases}$$

The parameter estimates are obtained by LS.

(a) Let $\mathbb{E}[U_t^2] = \sigma^2$. Determine the asymptotic covariance matrices of the estimation errors

$$\begin{cases} \mathcal{M}_1 : \quad \delta_1 = \frac{\sqrt{n}}{\lambda} \begin{bmatrix} \hat{a} \\ \hat{b} - b_0 \end{bmatrix} \\ \mathcal{M}_2 : \quad \delta_2 = \frac{\sqrt{n}}{\lambda} \begin{bmatrix} \hat{b}_1 - b \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix} \end{cases}$$

where n denotes the number of samples.

(b) Let the adequacy of a model structure be expressed by its ability to express the system's output one step ahead, when $\mathbb{E}[U_t^2] = s^2 \neq \sigma^2$. Then consider

$$A_{\mathcal{M}_1} = \mathbb{E}\left[\mathbb{E}[\epsilon_t^2(\hat{\theta}_{\mathcal{M}_1}) \mid \hat{\theta}_{\mathcal{M}_1}]\right]$$

and

$$A_{\mathcal{M}_2} = \mathbb{E}\left[\mathbb{E}[\epsilon_t^2(\hat{\theta}_{\mathcal{M}_2}) \mid \hat{\theta}_{\mathcal{M}_2}]\right]$$

Determine asymptotically (for $n \to \infty$) valid approximations for $A_{\mathcal{M}_1}$ and $A_{\mathcal{M}_2}$. Show that the inequality $A_{\mathcal{M}_1} \leq A_{\mathcal{M}_2}$ does not necessarily hold. Does this principle contradict the parsimony principle?

Solution: (a) The covariances of the estimates under $\mathcal{M}_1$ and $\mathcal{M}_2$ are given as

$$\begin{cases} P_1 = \begin{bmatrix} \frac{1}{\lambda^2 + \sigma^2 b^2} & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} \\ P_2 = \frac{1}{\sigma^2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{cases}$$

Note that the variance of the estimate of b equals $1/\sigma^2$ in both cases.

(b) Straightforward calculations give

$$A_{\mathcal{M}_1} = \mathbb{E}\left[\mathbb{E}[D_t + aY_{t-1} - (\hat{b} - b)U_{t-1}]^2\right] = \lambda^2 + \hat{a}^2(b^2 s^2 + \lambda^2) + (\hat{b} - b)^2 s^2,$$

equals

$$\mathbb{E}\left[\lambda^2 + \mathbb{E}\left[\begin{bmatrix}(\lambda^2 + s^2 b^2) & s^2\end{bmatrix}\begin{bmatrix}\hat{a}^2 \\ (\hat{b} - b)^2\end{bmatrix}\right]\right]$$

Similarly, one has that

$$A_{\mathcal{M}_2} = \mathbb{E}\left[\mathbb{E}[D_t - (\hat{b}_1 - b)U_{t-1} - \hat{b}_2 U_{t-2} - \hat{b}_3 U_{t-3}]^2\right] = \lambda^2 + s^2\mathbb{E}\left[(\hat{b}_1 - b) + \hat{b}_2^2 + \hat{b}_3^2\right]$$

Inserting the expressions for the estimates of $\mathcal{M}_1$ and $\mathcal{M}_2$ gives asymptotically valid expressions for $A_{\mathcal{M}_1}$ and $A_{\mathcal{M}_2}$ as

$$A_{\mathcal{M}_1} = \lambda^2 + \frac{\lambda^2}{n}\left(\frac{\lambda^2 + s^2 b^2}{\lambda^2 + \sigma^2 b^2} + \frac{s^2}{\sigma^2}\right)$$

and

$$A_{\mathcal{M}_2} = \lambda^2 + \frac{\lambda}{n}\frac{3s^2}{\sigma^2}$$

Note that for $s = 0$ we have that

$$A_{\mathcal{M}_1} \approx \lambda^2(1 + 2/n) < A_{\mathcal{M}_2} \approx \lambda^2(1 + 3/n)$$

For $s \neq \sigma$, one may however obtain the converse. For example take $b = \lambda^2 = \sigma^2 = 1$ and $s^2 = 0.1$. Thus the 'simpler' structure $\mathcal{M}_1$ may on the average lead to less accurate predictions than those obtained by $\mathcal{M}_2$! Note that since $\mathcal{M}_1 \not\subset \mathcal{M}_2$, this example is not in contradiction to the parsimony principle.