ID3
ARX Identification

3.14 ORDER ESTIMATION AND MODEL VALIDATION

In previous sections it has always been assumed to know the correct value of the model order n, and to use this value in estimating the parameters. In the identification of real processes the only available information concerns input–output sequences to be interpreted by an ARX model. In such a context the estimation of a suitable model order consists in performing on the data and/or on the model, tests allowing a comparison between different orders. Validating a model consists, instead, in evaluating its capability to describe the process that has generated the data in a way compatible with its planned use.

These problems, even if conceptually different, exhibit many synergies and interactions. A validation failure leads, in fact, to reconsider all identification steps, starting from the estimation of the order of the model; moreover order selection and validation tests are often based on common criteria.

3.14.1 PPCRE and singularity of the moments matrix

Consider an ARX process with order n, an integer $k > 0$ and the $(N \times 2k+1)$ matrix of input–output samples given by

$$H_k^* = \left[H_k(y)\ H_k(u)\ \ y_k^\circ \right] \tag{3.14.1))}$$

where $H_k(y)$ and $H_k(u)$ are defined by (3.3.6) and (3.3.7) and

$$y_k^\circ = \left[y(k+1)\ \ldots\ y(L) \right]^T; \tag{3.14.2}$$

so that $y_n^\circ = y^\circ$. Under identifiability conditions and for $e(t) = 0$ we would find

$$\text{rank } H_k^* = 2k+1 \qquad \text{for } k < n \tag{3.14.3a}$$

$$\text{rank } H_k^* = 2k \qquad \text{for } k = n. \tag{3.14.3b}$$

It would be then possible to consider the sequence of second order moments matrices

$$S_1 \; S_2 \; \ldots \; S_{n-1} \; S_n \; \ldots \tag{3.14.4}$$

where

$$S_k = H_k^{*T} H_k^* \tag{3.14.5}$$

evaluating their singularity; the position of the first singular matrix S_n, would define the correct order for the model. The presence of $e(t)$ in model (3.1.1) leads to the nonsingularity of every matrix in sequence (3.14.4), even for $k \geq n$. Consider now the following matrices

$$H_n^* = \left[H \; y^\circ \right] = \left[H \; (H\theta^* + e^\circ) \right] \tag{3.14.6}$$

$$S_n = H_n^{*T} H_n^* = \left[H \; y^\circ \right]^T \left[H \; y^\circ \right] = \begin{bmatrix} H^T \\ (H\theta^* + e^\circ)^T \end{bmatrix} \left[H \; (H\theta^* + e^\circ) \right] \tag{3.14.7}$$

$$= \begin{bmatrix} H^T H & H^T H\theta^* + H^T e^\circ \\ \theta^{*T} H^T H + e^{\circ T} H & \theta^{*T} H^T H\theta^* + \theta^{*T} H^T e^\circ + e^{\circ T} H\theta^* + e^{\circ T} e^\circ \end{bmatrix};$$

since $\mathrm{E}[H^T e^\circ] = 0$ (3.5.15), it follows that

$$\mathrm{E}\left[S_n \right] = \begin{bmatrix} \mathrm{E}[H^T H] & \mathrm{E}[H^T H]\theta^* \\ \theta^{*T} \mathrm{E}[H^T H] & \theta^{*T} \mathrm{E}[H^T H]\theta^* + \mathrm{E}[e^{\circ T} e^\circ] \end{bmatrix}. \tag{3.14.8}$$

Relation (3.14.8) shows that $e(\cdot)$ affects only the last element of the main diagonal of $\mathrm{E}[S_n]$ through the additional term $\mathrm{E}[e^{\circ T} e^\circ] = N\,\sigma_e^2$. Assume now to know $\mathrm{E}[S_n] > 0$; using a property of the determinant of partitioned matrices we can write

$$\begin{aligned} \det \mathrm{E}\left[S_n \right] &= \left(\theta^{*T} \mathrm{E}[H^T H]\,\theta^* + N\,\sigma_e^2 - \theta^{*T} \mathrm{E}[H^T H]\,\theta^* \right) \det \mathrm{E}\left[H^T H \right] \\ &= N\,\sigma_e^2 \det \mathrm{E}\left[H^T H \right] \end{aligned} \tag{3.14.9}$$

$$\sigma_e^2 = \frac{\det \mathrm{E}\left[S_n \right]}{N \det \mathrm{E}\left[H^T H \right]}. \tag{3.14.10}$$

Consider now the quantity

$$\sigma_{ek}^2 = \frac{\det \mathrm{E}\left[S_k \right]}{N \det \mathrm{E}\left[H_k^T H_k \right]} \tag{3.14.11}$$

where $H_k = [H_k(y)\; H_k(u)]$; $\sigma_{ek}^2 > \sigma_e^2$ for $k < n$ because $\mathrm{E}[S_k] > 0$ even for null equation errors, while for $k \geq n$, $\sigma_{ek}^2 \simeq \sigma_e^2$. A possible criterion to evaluate n can

thus be obtained by substituting S_k and $H_k^T H_k$ to their expected values in (3.14.11). If N is large enough we should observe a sequence of decreasing values followed by a stabilization once the correct value of the order is reached. This criterion can be used to evaluate a suitable order or, at least, an interval of admissible orders for the model before computing its parameters.

Previous results can be used to set up a singularity criterion for moments matrices useful in selecting models on the basis of their previsional properties even when the data have not been generated by an ARX process. Consider, to this purpose, the quantity

$$s_k = \frac{\det S_k}{N \det\left[H_k^T H_k\right]}; \tag{3.14.12}$$

it is easy to show that s_k coincides with the mean square prediction error $J(\theta)$ (3.3.2) for an ARX model with order k parameterized by $\theta_k^\circ = [H_k^T H_k\ ]^{-1} H_k^T y_k^\circ$. Assuming a null mean value for the residuals, the standard deviation of the prevision error is given by

$$\sigma_{\varepsilon k} = \sqrt{s_k} = \sqrt{\frac{\det S_k}{N \det\left[H_k^T H_k\right]}} \tag{3.14.13}$$

and, defining the percent one–step–ahead prediction error as 100 times the ratio between the standard deviation of the prediction error and that of the output, we obtain the *PPCRE* (Predicted PerCent Reconstruction Error) criterion given, for zero–mean output sequences, by

$$PPCRE(k) = 100 \sqrt{\frac{\det S_k}{y_k^{\circ T} y_k^\circ \det\left[H_k^T H_k\right]}}\,. \tag{3.14.14}$$

This criterion gives the prediction error of an order k ARX model without requiring any computation of its parameters. The application of the *PPCRE* criterion consists in computing the sequence

$$PPCRE(1)\ \ PPCRE(2)\ \ \ldots\ \ PPCRE(k)\ \ PPCRE(k+1)\ \ \ldots \tag{3.14.15}$$

and in selecting the minimal order that, once increased, does not lead to a significantly better performance. Of course sequence (3.14.15) has the same properties of (3.14.11) and it is thus reasonable to expect a stabilization when k reaches the order of the process. Previous results allow deducing a different expression for the estimate (3.10.5) of σ_e^2, given by

$$\hat{\sigma}_e^2 = \frac{1}{N-d} \sum_{t=n+1}^{L} \varepsilon(t)^2 = \frac{N}{N-d} J(\theta) = \frac{\det S_n}{(N-d) \det\left[H^T H\right]}. \tag{3.14.16}$$

Relation (3.14.16) gives the same estimate as (3.10.5) but does not require computing the parameters of the model and its prevision; it can also be used in the application of FPE, AIC and MDL criteria with the same advantage.

3.14.2 FPE (Final Prediction Error) criterion

This criterion consists in minimizing the expected value of the prediction error variance σ_ε^2 (3.10.3). Substituting estimate $\hat{\sigma}_\varepsilon^2$ (3.10.5) in (3.10.4) we obtain, for an order k model, the quantity

$$FPE(k) = \frac{N+d}{N-d} J\left(\theta_N^\circ\right) = \frac{N+d}{N(N-d)} \sum_{t=n+1}^{L} \varepsilon(t)^2 \tag{3.14.17}$$

that, differently from $J(\theta_N^\circ)$, tends to infinity for $d \to N$. Applying this criterion means computing the sequence

$$FPE(1) \quad FPE(2) \quad \ldots \quad FPE(k) \quad FPE(k+1) \quad \ldots \tag{3.14.18}$$

and selecting, as correct model order, the integer associated with the minimum value of the criterion. The analysis of the properties of this criterion, under the assumption that the data have been generated by an ARX process, shows that the probability of overestimating the order of the model is non null.

3.14.3 AIC (Akaike Information Criterion)

The FPE criterion penalizes, for increasing values of d, the decrease of $J(\theta_N^\circ)$. A family of alternative criteria can be based on structures of the type

$$N \log\left[J(\theta_N^\circ)\right] + \gamma(N,d) \tag{3.14.19}$$

where $\gamma(N,d)$ penalizes high order models. Choosing $\gamma(N,d) = 2d$ we obtain the AIC criterion

$$AIC = N \log\left[J(\theta_N^\circ)\right] + 2d \tag{3.14.20}$$

which is asymptotically equivalent to FPE since, for large values of N, $\log FPE$ is given by

$$\begin{aligned} \log\left[\frac{1+d/N}{1-d/N} J(\theta_N^\circ)\right] &= \log\left[1+\frac{d}{N}\right] - \log\left[1-\frac{d}{N}\right] + \log\left[J(\theta_N^\circ)\right] \\ &\simeq \frac{2d}{N} + \log\left[J(\theta_N^\circ)\right]. \end{aligned} \tag{3.14.21}$$

Note that the AIC criterion reduces the penalization on the order of the model for high values of N, leading thus to select models with larger orders when N is high.

3.14.4 MDL (Minimum Description Length) criterion

While FPE and AIC criteria are based on statistical considerations, the MDL criterion is based on the minimization of the information necessary to describe a model and its prediction error. Increasing the dimension of the model, an increased information is necessary to describe its parameters while the information required to describe its prediction error decreases. The MDL criterion consists in the expression

$$MDL = \log\big[N\big]d + N\log\big[J(\theta_N^\circ)\big]. \tag{3.14.22}$$

The only difference with AIC consists in multiplying the number of parameters, d, by $\log[N]$ instead than by 2. For usual lengths of the sequences used for identification (some hundred samples) $\log[N] > 2$ so that the MDL criterion leads to models with orders lower than those selected by AIC and FPE.

3.14.5 Whiteness test on residuals

It has been shown that the prevision errors (residuals) of an ARX model with correct order estimated from input–output sequences generated by an ARX process, constitute asymptotically, a white process. The whiteness of residuals is usually evaluated by computing the sample covariances

$$R_\varepsilon^N(\tau) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon(t)\,\varepsilon(t+\tau) \qquad (\tau = 1, \ldots, M). \tag{3.14.23}$$

If $\varepsilon(t)$ is a white process, then the quantity

$$\zeta_{N,M} = \frac{N}{R_\varepsilon(0)^2}\sum_{\tau=1}^{M}\big(R_\varepsilon^N(\tau)\big)^2 \tag{3.14.24}$$

will have, asymptotically, a $\chi^2(M)$ distribution. The independence between residuals can thus be verified testing whether $\zeta_{N,M} < \chi_\alpha^2(M)$, the α level of $\chi^2(M)$ distribution, for a significant choice of α. Typical choices range from 0.05 to 0.005.

3.14.6 Test on the independence between residuals and previous inputs

The whiteness of residuals should lead also to the independence between $\varepsilon(t)$ and the input sequence $u(t)$. This independence can be evaluated computing the sample covariances

$$R_{\varepsilon u}^N(\tau) = \frac{1}{N}\sum_{t=1}^{N}u(t)\varepsilon(t+\tau). \tag{3.14.25}$$

When $\varepsilon(\cdot)$ and $u(\cdot)$ are independent, the variable $\sqrt{N}R_{\varepsilon u}^N(\tau)$ assumes a Gaussian distribution with null expected value and variance $\sigma_{\varepsilon u}^2$ given by

$$\sigma_{\varepsilon u}^2 = \sum_{k=-\infty}^{\infty}R_\varepsilon(k)\,R_u(k) \tag{3.14.26}$$

where

$$R_\varepsilon(k) = \mathrm{E}\left[\,\varepsilon(t)\varepsilon(t-k)\,\right] \tag{3.14.27a}$$

$$R_u(k) = \mathrm{E}\left[\,u(t)u(t-k)\,\right]. \tag{3.14.27b}$$

An efficient way of reporting the results of this test consists in plotting $R^N_{\varepsilon u}(\tau)$ against τ; since $\sigma^2_{\varepsilon u}$ does not depend on τ, the confidence limits are represented by horizontal lines and any deviation from non correlation assumptions can be easily appreciated. This test can be used also to check the presence of delays in the process and/or the choice of improper values for the delays inserted in the model.

The correlation between $\varepsilon(t)$ and $u(t-\tau)$ can evidentiate also other inadequacies in the description of a process by means of a model; the scattering of the points $(\varepsilon(t), u(t-\tau))$ has been, for instance, proposed as evidence of the presence of nonlinear behaviors not described by the model. Of course every practical implementation of this test will approximate (3.14.26) with a limited number of terms.

3.14.7 Cross validation by simulation

A very effective way of evaluating the adequacy and flexibility of identified models consists in their use for performing complete simulations (i.e. using only the initial samples of the observed outputs) and in comparing the obtained previsions with observed output samples. This procedure, that can be applied when a single set of data is available, gives best results when applied to sequences different from those used to identify the model. A quadratic criterion like (3.3.2) applied to the difference between the observed output and that obtained by simulation can then be used to compare models with different orders.

A possible variation of this criterion consists in performing, when suggested by specific applications of the models, k–step–ahead simulations, that is using, at time t, only the observed values of the output at times lower than $t-k$ and those obtained by simulation at times $t-k, \ldots, t-1$. Obviously a hybrid simulation of this kind coincides with a complete simulation when $k > n$.