7

PARAMETER ESTIMATION METHODS

Suppose a set of candidate models has been selected, and it is parametrized as a model structure (see Sections 4.5 and 5.7), using a parameter vector θ. The search for the best model within the set then becomes a problem of determining or estimating θ. There are many different ways of organizing such a search and also different views on what one should search for. In the present chapter we shall concentrate on the latter aspect: what should be meant by a "good model"? Computational issues (i.e., how to organize the actual search) will be dealt with in Chapters 10 and 11. The evaluation of the properties of the models that result under various conditions and using different methods is carried out in Chapters 8 and 9. In Chapter 15 we return to the estimation methods, and give a more user-oriented summary of recommended procedures.

7.1 GUIDING PRINCIPLES BEHIND PARAMETER ESTIMATION METHODS

Parameter Estimation Methods

We are now in the situation that we have selected a certain model structure $\mathcal{M}$, with particular models $\mathcal{M}(\theta)$ parametrized using the parameter vector $\theta \in D_{\mathcal{M}} \subset \mathbf{R}^d$. The set of models thus defined is

$$\mathcal{M}^* = \{\mathcal{M}(\theta) | \theta \in D_{\mathcal{M}}\} \tag{7.1}$$

Recall that each model represents a way of predicting future outputs. The predictor could be a linear filter, as discussed in Chapter 4:

$$\mathcal{M}(\theta) : \hat{y}(t|\theta) = W_y(q, \theta)y(t) + W_u(q, \theta)u(t) \tag{7.2}$$

This could correspond to one-step-ahead prediction for an underlying system description

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{7.3}$$

in which case

$$W_y(q, \theta) = \left[1 - H^{-1}(q, \theta)\right], \qquad W_u(q, \theta) = H^{-1}(q, \theta)G(q, \theta) \tag{7.4}$$

but it could also be arrived at from other considerations.

The predictor could also be a nonlinear filter, as discussed in Chapter 5, in which case we write it as a general function of past data Z^{t-1}:

$$\mathcal{M}(\theta) : \hat{y}(t|\theta) = g(t, Z^{t-1}; \theta) \tag{7.5}$$

The model $\mathcal{M}(\theta)$ may also contain (model) assumptions about the character of the associated prediction errors, such as their variances ($\lambda(\theta)$) or their probability distribution (PDF $f_e(x, \theta)$).

We are also in the situation that we have collected, or are about to collect, a batch of data from the system:

$$Z^N = [y(1), u(1), y(2), u(2), \ldots, y(N), u(N)] \tag{7.6}$$

The problem we are faced with is to decide upon how to use the information contained in Z^N to select a proper value $\hat{\theta}_N$ of the parameter vector, and hence a proper member $\mathcal{M}(\hat{\theta}_N)$ in the set $\mathcal{M}^*$. Formally speaking, we have to determine a mapping from the data Z^N to the set $D_{\mathcal{M}}$:

$$Z^N \rightarrow \hat{\theta}_N \in D_{\mathcal{M}} \tag{7.7}$$

Such a mapping is a *parameter estimation method.*

Evaluating the Candidate Models

We are looking for a test by which the different models' ability to "describe" the observed data can be evaluated. We have stressed that the essence of a model is its prediction aspect, and we shall also judge its performance in this respect. Thus let the prediction error given by a certain model $\mathcal{M}(\theta_*)$ be given by

$$\varepsilon(t, \theta_*) = y(t) - \hat{y}(t|\theta_*) \tag{7.8}$$

When the data set Z^N is known, these errors can be computed for $t = 1, 2, \ldots, N$.

A "good" model, we say, is one that is good at predicting, that is, one that produces small prediction errors when applied to the observed data. Note that there is considerable flexibility in selecting various predictor functions, and this gives a corresponding freedom in defining "good" models in terms of prediction performance. A guiding principle for parameter estimation thus is:

> *Based on Z^t we can compute the prediction error $\varepsilon(t, \theta)$ using (7.8). At time $t = N$, select $\hat{\theta}_N$ so that the prediction errors $\varepsilon(t, \hat{\theta}_N)$, $t = 1, 2, \ldots, N$, become as small as possible.* (7.9)

The question is how to qualify what "small" should mean. In this chapter we shall describe two such approaches. One is to form a scalar-valued norm or criterion function that measures the size of ε. This approach is dealt with in Sections 7.2 to 7.4. Another approach is to demand that $\varepsilon(t, \hat{\theta}_N)$ be uncorrelated with a given data sequence. This corresponds to requiring that certain "projections" of $\varepsilon(t, \hat{\theta}_N)$ are zero and is further discussed in Sections 7.5 and 7.6.

7.2 MINIMIZING PREDICTION ERRORS

The prediction-error sequence in (7.8) can be seen as a vector in $\mathbf{R}^N$. The "size" of this vector could be measured using any norm in $\mathbf{R}^N$, quadratic or nonquadratic. This leaves a substantial amount of choices. We shall restrict the freedom somewhat by only considering the following way of evaluating "how large" the prediction-error sequence is: Let the prediction-error sequence be filtered through a stable linear filter $L(q)$:

$$\varepsilon_F(t, \theta) = L(q)\varepsilon(t, \theta), \qquad 1 \le t \le N \tag{7.10}$$

Then use the following norm:

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \ell(\varepsilon_F(t, \theta)) \tag{7.11}$$

where $\ell(\cdot)$ is a scalar-valued (typically positive) function.

The function $V_N(\theta, Z^N)$ is, for given Z^N, a well-defined scalar-valued function of the model parameter θ. It is a natural measure of the validity of the model $\mathcal{M}(\theta)$. The estimate $\hat{\theta}_N$ is then defined by minimization of (7.11):

$$\hat{\theta}_N = \hat{\theta}_N(Z^N) = \arg\min_{\theta \in D_{\mathcal{M}}} V_N(\theta, Z^N) \tag{7.12}$$

Here arg min means "the minimizing argument of the function." If the minimum is not unique, we let arg min denote the set of minimizing arguments. The mapping (7.7) is thus defined implicitly by (7.12).

This way of estimating θ contains many well-known and much used procedures. We shall use the general term *prediction-error identification methods* (PEM) for the family of approaches that corresponds to (7.12). Particular methods, with specific "names" attached to themselves, are obtained as special cases of (7.12), depending on the choice of $\ell(\cdot)$, the choice of prefilter $L(\cdot)$, the choice of model structure, and, in some cases, the choice of method by which the minimization is realized. We shall give particular attention to two especially well known members in the family (7.12) in the subsequent two sections. First, however, let us discuss some aspects on the choices of $L(q)$ and $\ell(\cdot)$ in (7.10) and (7.11). See also Section 15.2.

Choice of L

The effect of the filter L is to allow extra freedom in dealing with non-momentary properties of the prediction errors. Clearly, if the predictor is linear and time invariant, and y and u are scalars, then the result of filtering ε, is the same as first filtering the input-output data and then applying the predictors.

The effect of L is best understood in a frequency-domain interpretation and a full discussion will be postponed to Section 14.4. It is clear, however, that by the use of L, effects of high-frequency disturbances, not essential to the modeling problem, or slow drift terms and the like, can be removed. It also seems reasonable that certain properties of the models may be enhanced or suppressed by a properly selected L. L thus acts like *frequency weighting*.

The following particular aspect of the filtering (7.10) should be noted. If a model (7.3) is used, the filtered error $\varepsilon_F(t, \theta)$ is given by

$$\varepsilon_F(t, \theta) = L(q)\varepsilon(t, \theta) = \left[L^{-1}(q)H(q, \theta)\right]^{-1}[y(t) - G(q, \theta)u(t)] \quad (7.13)$$

The effect of prefiltering is thus identical to changing the noise model from $H(q, \theta)$ *to*

$$\overline{H}_L(q, \theta) = L^{-1}(q)H(q, \theta) \quad (7.14)$$

When we describe and analyze methods that employ general noise models in linear systems, we shall usually confine ourselves to $L(q) \equiv 1$, since the option of prefiltering is taken care of by the freedom in selecting $H(q, \theta)$. A discussion of the use and effects of $L(q)$ in practical terms will be given in Section 14.4.

Choice of ℓ

For the choice of $\ell(\cdot)$, a first candidate would be a quadratic norm:

$$\ell(\varepsilon) = \tfrac{1}{2}\varepsilon^2 \quad (7.15)$$

and this is indeed a standard choice, which is convenient both for computation and analysis. Questions of robustness against bad data may, however, warrant other norms, which we shall discuss in some detail in Section 15.2. One may also conceive situations where the "best" norm is not known beforehand so that it is reasonable to parametrize the norm itself:

$$\ell(\varepsilon, \theta) \quad (7.16)$$

Often the parametrization of the norm is independent of the model parametrization:

$$\theta = \begin{bmatrix} \theta' \\ \alpha \end{bmatrix}: \ \ell(\varepsilon(t, \theta), \theta) = \ell\left(\varepsilon(t, \theta'), \alpha\right) \quad (7.17)$$

An exception to this case is given in Problem 7E.4.

Time-varying Norms

It may happen that measurements at different time instants are considered to be of varying reliability. The reason may be that the degree of noise corruption changes or that certain measurements are less representative for the system's properties. In such cases we are motivated to let the norm ℓ be time varying:

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \ell(\varepsilon(t, \theta), \theta, t) \quad (7.18)$$

In this way less reliable measurements can be associated with less weight in the criterion.

We shall frequently work with a criterion where the weighting is made explicitly by a weighting function $\beta(N, t)$:

$$V_N(\theta, Z^N) = \sum_{t=1}^{N} \beta(N, t)\ell(\varepsilon(t, \theta), \theta) \quad (7.19)$$

For fixed N, the N-dependence of $\beta(N, t)$ is of course immaterial. However, when estimates $\hat{\theta}_N$ for different N are compared, as for example in recursive identification (see Chapter 11), it becomes interesting to discuss how $\beta(N, t)$ varies with N. We shall return to this issue in Section 11.2.

Frequency-domain Interpretation of Quadratic Prediction-error Criteria for Linear Time-invariant Models

Let us consider the quadratic criterion error (7.12) and (7.15) for the standard linear model (7.3)

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \tfrac{1}{2}\varepsilon^2(t, \theta) \tag{7.20}$$

$$\varepsilon(t, \theta) = H^{-1}(q, \theta)[y(t) - G(q, \theta)u(t)]$$

Let $E_N(2\pi k/N, \theta)$, $k = 0, 1, \ldots, N-1$, be the DFT of $\varepsilon(t, \theta)$, $t = 1, 2, \ldots, N$:

$$E_N(2\pi k/N, \theta) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N} \varepsilon(t, \theta)e^{-2\pi ikt/N}$$

Then, by Parseval's relation (2.44),

$$V_N(\theta, Z^N) = \frac{1}{N}\frac{1}{2}\sum_{k=1}^{N-1} |E_N(2\pi k/N, \theta)|^2 \tag{7.21}$$

Now let

$$w(t, \theta) = G(q, \theta)u(t)$$

Then the DFT of $w(t, \theta)$ is, according to Theorem 2.1,

$$W_N(\omega, \theta) = G(e^{i\omega}, \theta)U_N(\omega) + R_N(\omega)$$

with

$$|R_N(\omega)| \leq \frac{C}{\sqrt{N}}$$

The DFT of $s(t, \theta) = y(t) - w(t, \theta)$ then is

$$S_N(\omega, \theta) = Y_N(\omega) - G(e^{i\omega}, \theta)U_N(\omega) - R_N(\omega)$$

Finally,

$$\varepsilon(t, \theta) = H^{-1}(q, \theta)s(t, \theta)$$

has the DFT, again using Theorem 2.1,

$$E_N(\omega) = H^{-1}(e^{i\omega}, \theta)S_N(\omega, \theta) + \tilde{R}_N(\omega)$$

with

$$\left|\tilde{R}_N(\omega)\right| \leq \frac{C}{\sqrt{N}}$$

Inserting this into (7.21) gives

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{k=0}^{N-1}\frac{1}{2}\left|H(e^{2\pi ik/N}, \theta)\right|^{-2} \times \left|Y_N(2\pi k/N) - G(e^{2\pi ik/N}, \theta)U_N(2\pi k/N)\right|^2 + \overline{R}_N$$

with $|\overline{R}_N| \leq C/\sqrt{N}$, or, using the definition of the ETFE $\hat{\hat{G}}_N$ in (6.24),

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{k=0}^{N-1}\left\{\frac{1}{2}\left|\hat{\hat{G}}_N(e^{2\pi ik/N}) - G(e^{2\pi ik/N}, \theta)\right|^2 \times Q_N(2\pi k/N, \theta) + \overline{R}_N\right\} \tag{7.22}$$

with

$$Q_N(\omega, \theta) = \frac{|U_N(\omega)|^2}{\left|H(e^{i\omega}, \theta)\right|^2} \tag{7.23}$$

First notice that, apart from the remainder term $\overline{R}_N$, the expression (7.22) coincides with the weighted least-squares criterion for a model:

$$\hat{\hat{G}}_N(e^{2\pi ik/N}) = G(e^{2\pi ik/N}, \theta) + v(k) \tag{7.24}$$

Compare with (II.96) and (II.97). According to Lemma 6.1, the variance of $v(k)$ is, asymptotically, $\Phi_v(2\pi k/N)/|U_N(2\pi k/N)|^2$, so the weighting coefficient $Q_N(\omega, \theta)$ is the inverse variance, which is optimal for linear regressions, according to (II.65). In (7.23) the unknown noise spectrum $\Phi_v(\omega)$ is replaced by the model noise spectrum $\left|H(e^{i\omega}, \theta)\right|^2$. Consequently, the prediction-error methods can be seen as methods of fitting the ETFE to the model transfer function with a weighted norm, corresponding to the model signal-to-noise ratio at the frequency in question. For notational reasons, it is instructive to rewrite the sum (7.22) approximately as an integral:

$$V_N(\theta, Z^N) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\left|\hat{\hat{G}}_N(e^{i\omega}) - G(e^{i\omega}, \theta)\right|^2 Q_N(\omega, \theta)\, d\omega \tag{7.25}$$

The shift of integration interval from $(0, 2\pi)$ to $(-\pi, \pi)$ is possible since the integrand is periodic.

With this interpretation we have described the prediction-error estimate as an alternative way of smoothing the ETFE, showing a strong conceptual relationship to the spectral analysis methods of Section 6.4. See Problem 7G.2 for a direct tie.

When we specialize to the case of a time series [no input and $G(q, \theta) \equiv 0$], the criterion (7.25) takes the form

$$V_N(\theta, Z^N) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{Y_N(\omega)}{H(e^{i\omega}, \theta)} \right|^2 d\omega \tag{7.26}$$

Such parametric estimators of spectra are known as "Whittle-type estimators," after Whittle (1951).

In Section 7.8 we shall return to frequency domain criteria. There, however, we take another viewpoint and assume that the observed data are in the frequency domain, being Fourier transforms of the input and output time domain signals.

Multivariable Systems (*)

For multioutput systems, the counterpart of the quadratic criterion is

$$\ell(\varepsilon) = \tfrac{1}{2}\varepsilon^T \Lambda^{-1} \varepsilon \tag{7.27}$$

for some symmetric, positive semidefinite $p \times p$ matrix Λ that weights together the relative importance of the components of ε.

One might discuss what is the best choice of norm Λ. We shall do that in some detail in Section 15.2. Here we only remark that, just as in (7.16), the parameter vector θ could be extended to include components of Λ, and the function ℓ will then be an appropriate function of θ.

As a variant of the criterion (7.11), where a scalar $\ell(\varepsilon)$ is formed for each t, we could first form the $p \times p$ matrix

$$Q_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t, \theta)\varepsilon^T(t, \theta) \tag{7.28}$$

and let the criterion be a scalar-valued function of this matrix:

$$V_N(\theta, Z^N) = h\left(Q_N(\theta, Z^N)\right) \tag{7.29}$$

The criterion (7.27) is then obtained by

$$h(Q) = \tfrac{1}{2}\mathrm{tr}(Q\Lambda^{-1}) \tag{7.30}$$

7.3 LINEAR REGRESSIONS AND THE LEAST-SQUARES METHOD

Linear Regressions

We found in both Sections 4.2 and 5.2 that linear regression model structures are very useful in describing basic linear and nonlinear systems. The linear regression employs a predictor (5.67)

$$\hat{y}(t|\theta) = \varphi^T(t)\theta + \mu(t) \tag{7.31}$$

that is linear in θ. Here φ is the vector of regressors, the *regression vector*. Recall that for the ARX structure (4.7) we have

$$\varphi(t) = [-y(t-1) \ -y(t-2) \dots -y(t-n_a) \quad u(t-1) \dots u(t-n_b)]^T \tag{7.32}$$

In (7.31), $\mu(t)$ is a known data-dependent vector. For notational simplicity we shall take $\mu(t) = 0$ in the remainder of this section: it is quite straightforward to include it. See Problem 7D.1.

Linear regression forms a standard topic in statistics. The reader could consult Appendix II for a refresher of basic properties. The present section can, however, be read independently of Appendix II.

Least-squares Criterion

With (7.31) the prediction error becomes

$$\varepsilon(t, \theta) = y(t) - \varphi^T(t)\theta$$

and the criterion function resulting from (7.10) and (7.11), with $L(q) = 1$ and $\ell(\varepsilon) = \frac{1}{2}\varepsilon^2$, is

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \tfrac{1}{2}\left[y(t) - \varphi^T(t)\theta\right]^2 \tag{7.33}$$

This is the *least-squares criterion* for the linear regression (7.31). The unique feature of this criterion, developed from the linear parametrization and the quadratic criterion, is that it is a quadratic function in θ. Therefore, it can be minimized analytically, which gives, provided the indicated inverse exists,

$$\hat{\theta}_N^{\text{LS}} = \arg\min V_N(\theta, Z^N) = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) \tag{7.34}$$

the *least-squares estimate (LSE)* (see Problem 7D.2).

Introduce the $d \times d$ matrix

$$R(N) = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t) \tag{7.35}$$

and the d-dimensional column vector

$$f(N) = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) \tag{7.36}$$

In the case (7.32), $\varphi(t)$ contains lagged input and output variables, and the entries of the quantities (7.35) and (7.36) will be of the form

$$[R(N)]_{ij} = \frac{1}{N}\sum_{t=1}^{N} y(t-i)y(t-j), \qquad 1 \le i, j \le n_a$$

and similar sums of $u(t-r)\cdot u(t-s)$ or $u(t-r)\cdot y(t-s)$ for the other entries of $R(N)$. That is, they will consist of estimates of the covariance functions of $\{y(t)\}$ and $\{u(t)\}$. The LSE can thus be computed using only such estimates and is therefore related to correlation analysis, as described in Section 6.1.

Properties of the LSE

The least-squares method is a special case of the prediction-error identification method (7.12). An analysis of its properties is therefore contained in the general treatment in Chapters 8 and 9. It is, however, useful to include a heuristic investigation of the LSE at this point.

Suppose that the observed data actually have been generated by

$$y(t) = \varphi^T(t)\theta_0 + v_0(t) \tag{7.37}$$

for some sequence $\{v_0(t)\}$. We may think of θ_0 as a "true value" of the parameter vector.

As in (1.14)–(1.15) we find that

$$\lim_{N\to\infty} \hat{\theta}_N^{\mathrm{LS}} - \theta_0 = \lim_{N\to\infty} R^{-1}(N)\frac{1}{N}\sum_{t=1}^{N}\varphi(t)v_0(t) = (R^*)^{-1}f^*.$$

$$R^* = \bar{E}\varphi(t)\varphi^T(t), \quad f^* = \bar{E}\varphi(t)v_0(t) \tag{7.38}$$

provided v_0 and φ are quasi-stationary, so that Theorem 2.3 can be applied. For the LSE to be *consistent*, that is, for $\hat{\theta}_N^{\mathrm{LS}}$ to converge to θ_0, we thus have to require:

i. R^* is non-singular. This will be secured by the input properties, as in (1.17)–(1.18), and discussed in much more detail in Chapter 13.

ii. $f^* = 0$. This will be the case if either:

(a) $\{v_0(t)\}$ is a sequence of independent random variables with zero mean values (white noise). Then $v_0(t)$ will not depend on what happened up to time $t-1$ and hence $E\varphi(t)v_0(t) = 0$.

(b) The input sequence $\{u(t)\}$ is independent of the zero mean sequence $\{v_0(t)\}$ and $n_a = 0$ in (7.32). Then $\varphi(t)$ contains only u-terms and hence $E\varphi(t)v_0(t) = 0$.

When $n_a > 0$ so that $\varphi(t)$ contains $y(k)$, $t - n_a \le k \le t-1$, and $v_0(t)$ is not white noise, then (usually) $E\varphi(t)v_0(t) \neq 0$. This follows since $\varphi(t)$ contains $y(t-1)$, while $y(t-1)$ contains the term $v_0(t-1)$ that is correlated with $v_0(t)$. Therefore, we may expect consistency only in cases (a) and (b).

Weighted Least Squares

Just as in (7.18) and (7.19), the different measurements could be assigned different weights in the least-squares criterion:

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \alpha_t \left[y(t) - \varphi^T(t)\theta\right]^2 \tag{7.39}$$

or

$$V_N(\theta, Z^N) = \sum_{t=1}^{N} \beta(N, t)\left[y(t) - \varphi^T(t)\theta\right]^2 \tag{7.40}$$

The expression for the resulting estimate is quite analogous to (7.34):

$$\hat{\theta}_N^{\mathrm{LS}} = \left[\sum_{t=1}^{N} \beta(N, t)\varphi(t)\varphi^T(t)\right]^{-1} \sum_{t=1}^{N} \beta(N, t)\varphi(t)y(t) \tag{7.41}$$

Multivariable Case (*)

If the output $y(t)$ is a p-vector and the norm (7.27) is used, the LS criterion takes the form

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \frac{1}{2}\left[y(t) - \boldsymbol{\varphi}^T(t)\theta\right]^T \Lambda^{-1}\left[y(t) - \boldsymbol{\varphi}^T(t)\theta\right] \tag{7.42}$$

This gives the estimate

$$\hat{\theta}_N^{\mathrm{LS}} = \left[\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\varphi}(t)\Lambda^{-1}\boldsymbol{\varphi}^T(t)\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\varphi}(t)\Lambda^{-1}y(t) \tag{7.43}$$

In case we use the particular parametrization (4.56) with $\boldsymbol{\theta}$ as an $r \times p$ matrix,

$$\hat{y}(t|\theta) = \boldsymbol{\theta}^T\varphi(t) \tag{7.44}$$

the LS criterion becomes

$$V_N(\boldsymbol{\theta}, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \left\|y(t) - \boldsymbol{\theta}^T\varphi(t)\right\|^2 \tag{7.45}$$

with the estimate

$$\hat{\boldsymbol{\theta}}_N^{\mathrm{LS}} = \left[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \varphi(t)y^T(t) \tag{7.46}$$

(see problem 7D.2). The expression (7.46) brings out the advantages of the structure (7.44): To determine the $r \times p$ estimate $\hat{\boldsymbol{\theta}}_N$, it is sufficient to invert an $r \times r$ matrix. In (7.43) θ is a $p \cdot r$ vector and the matrix inversion involves a $pr \times pr$ matrix.

Colored Equation-error Noise (*)

The LS method has many advantages, the most important one being that the global minimum of (7.33) can be found efficiently and unambiguously (no local minima other than global ones exist). Its main shortcoming relates to the asymptotic properties quoted previously: If, in a difference equation,

$$\begin{aligned} y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) \\ = b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) + v(t) \end{aligned} \tag{7.47}$$

the equation error $v(t)$ is not white noise, then the LSE will not converge to the true values of a_i and b_i. To deal with this problem, we may incorporate further modeling of the equation error $v(t)$ as discussed in Section 4.2, let us say

$$v(t) = \kappa(q)e(t) \tag{7.48}$$

with e white and κ linear filter. Models employing (7.48) will typically take us out from the LS environment, except in two cases, which we now discuss.

Known noise properties: If in (7.47) and (7.48) a_i and b_i are unknown, but κ is a known filter (not too realistic a situation), we have

$$A(q)y(t) = B(q)u(t) + \kappa(q)e(t) \tag{7.49}$$

Filtering (7.49) through the filter $\kappa^{-1}(q)$ gives

$$A(q)y_F(t) = B(q)u_F(t) + e(t) \tag{7.50}$$

where

$$y_F(t) = \kappa^{-1}(q)y(t), \qquad u_F(t) = \kappa^{-1}(q)u(t) \tag{7.51}$$

Since e is white, the LS method can be applied to (7.50) without problems. Notice that this is equivalent to applying the filter $L(q) = \kappa^{-1}(q)$ in (7.10).

High-order models: Suppose that the noise v can be well described by $\kappa(q) = 1/D(q)$ in (7.48), where $D(q)$ is a polynomial of degree r. [That is, $v(t)$ is supposed to be an autoregressive (AR) process of order r.] This gives

$$A(q)y(t) = B(q)u(t) + \frac{1}{D(q)}e(t) \tag{7.52}$$

or

$$A(q)D(q)y(t) = B(q)D(q)u(t) + e(t) \tag{7.53}$$

Applying the LS method to (7.53) with orders $n_A = n_a + r$ and $n_B = n_b + r$ gives, since e is white, consistent estimates of AD and BD. Hence the transfer function from u to y,

$$\frac{B(q)D(q)}{A(q)D(q)} = \frac{B(q)}{A(q)}$$

is correctly estimated. This approach was called *repeated least squares* in Åström and Eykhoff (1971). See also Söderström (1975b)and Stoica (1976).

Estimating State Space Models Using Least Squares Techniques (Subspace Methods)

A linear system can always be represented in state space form as in (4.84):

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + w(t) \\ y(t) &= Cx(t) + Du(t) + v(t) \end{aligned} \tag{7.54}$$

with white noises w and v. Alternatively we could just represent the input-output dynamics as in (4.80):

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) + v(t) \end{aligned} \tag{7.55}$$

where the noise at the output, v, very well could be colored. It should be noted that the input-output dynamics could be represented with a lower order model in (7.55) than in (7.54) since describing the noise character might require some extra states.

To estimate such a model, the matrices can be parameterized in ways that are described in Section 4.3 or Appendix 4A—either from physical grounds or as black boxes in canonical forms. Then these parameters can be estimated using the techniques dealt with in Section 7.4.

However, there are also other possibilities: We assume that we have no insight into the particular structure, and we would just estimate any matrices A, B, C, and D that give a good description of the input-output behavior of the system. Since there are an infinite number of such matrices that describe the same system (the similarity transforms), we will have to fix the coordinate basis of the state-space realization.

Let us for a moment assume that not only are u and y measured, but also the sequence of state vectors x. This would, by the way, fix the state-space realization coordinate basis. Now, with known u, y and x, the model (7.54) becomes a linear regression: the unknown parameters, all of the matrix entries in all the matrices, mix with measured signals in linear combinations. To see this clearly, let

$$Y(t) = \begin{bmatrix} x(t+1) \\ y(t) \end{bmatrix}, \quad \Theta = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

$$\Phi(t) = \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \quad E(t) = \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}$$

Then, (7.54) can be rewritten as

$$Y(t) = \Theta\Phi(t) + E(t) \tag{7.56}$$

From this, all the matrix elements in Θ can be estimated by the simple least squares method (which in the case of Gaussian noise and known covariance matrix coincides with the maximum likelihood method), as described above in (7.44)–(7.46). The covariance matrix for $E(t)$ can also be estimated easily as the sample sum of the squared model residuals. That will give the covariance matrices as well as the cross covariance matrix for w and v. These matrices will, among other things, allow us to compute the Kalman filter for (7.54). Note that all of the above holds without changes for multivariable systems, i.e., when the output and input signals are vectors.

The problem is how to obtain the state vector sequence x. Some basic realization theory was reviewed in Appendix 4A, from which the essential results can be quoted as follows:

Let a system be given by the impulse response representation

$$y(t) = \sum_{j=0}^{\infty} [h_u(j)u(t-j) + h_e(j)e(t-j)] \tag{7.57}$$

where u is the input and e the innovations. Let the formal k-step ahead predictors be defined by just deleting the contributions to $y(t)$ from $e(j), u(j)$: $j = t, \ldots, t-k+1$:

$$\hat{y}(t|t-k) = \sum_{j=k}^{\infty} [h_u(j)u(t-j) + h_e(j)e(t-j)] \tag{7.58}$$

No attempt is thus made to predict the inputs $u(j)$: $j = t, \ldots, t-k+1$ from past data. Define

$$\hat{Y}_r(t) = \begin{bmatrix} \hat{y}(t|t-1) \\ \vdots \\ \hat{y}(t+r-1|t-1) \end{bmatrix} \tag{7.59a}$$

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_r(1) & \ldots & \hat{Y}_r(N) \end{bmatrix} \tag{7.59b}$$

Then the following is true as $N \to \infty$ (see Lemmas 4A.1 and 4A.2 and their proofs):

1. The system (7.57) has an nth order minimal state space description if and only if the rank $\hat{\mathbf{Y}}$ is equal to n for all $r \geq n$.

2. The state vector of any minimal realization in innovations form can be chosen as linear combinations of $\hat{Y}_r$ that form a row basis for $\hat{\mathbf{Y}}$, i.e.,

$$x(t) = L\hat{Y}_r(t) \tag{7.60}$$

where the $n \times pr$ matrix L is such that $L\hat{\mathbf{Y}}$ spans $\hat{\mathbf{Y}}$. (p is the dimension of the output vector $y(t)$.)

Note that the canonical state space representations described in Appendix 4A correspond to L matrices that just pick out certain rows of $\hat{Y}_n$. In general, we are not confined to such choices, but may pick L so that $x(t)$ becomes a well-conditioned basis.

It is clear that the facts above will allow us to find a suitable state vector from data. The only remaining problem is to estimate the k-step ahead predictors. The true predictor $\hat{y}(t + k - 1|t - 1)$ is given by (7.58). The innovation $e(j)$ can be written as a linear combination of past input-output data. The predictor can thus be expressed as a linear function of $u(i)$, $y(i)$, $i \leq t - 1$. For practical reasons the predictor is approximated so that it only depends on a fixed and finite amount of past data, like the s_1 past outputs and the s_2 past inputs. This means that it takes the form

$$\begin{aligned}\hat{y}(t + k - 1|t - 1) = {} & \alpha_1 y(t - 1) + \ldots + \alpha_{s_1} y(t - s_1) \\ & + \beta_1 u(t - 1) + \ldots + \beta_{s_2} u(t - s_2)\end{aligned} \tag{7.61}$$

This predictor can then efficiently be determined by another linear least squares projection directly on the input output data. That is, set up the model

$$y(t + k - 1) = \theta_k^T \varphi_s(t) + \gamma_k^T U_\ell(t) + \varepsilon(t + k - 1) \tag{7.62}$$

or, dealing with all r predictors simultaneously

$$Y_r(t) = \Theta\varphi_s(t) + \Gamma U_\ell(t) + E(t) \tag{7.63}$$

Here:

$$\varphi_s(t) = \begin{bmatrix} y^T(t - 1) & \ldots & y^T(t - s_1) & u^T(t - 1) & \ldots & u^T(t - s_2) \end{bmatrix}^T \tag{7.64a}$$

$$U_\ell(t) = \begin{bmatrix} u^T(t) & \ldots & u^T(t + \ell - 1) \end{bmatrix}^T \tag{7.64b}$$

$$Y_r(t) = \begin{bmatrix} y^T(t) & \ldots & y^T(t + r - 1) \end{bmatrix}^T \tag{7.64c}$$

$$\Theta = \begin{bmatrix} \theta_1 & \ldots & \theta_r \end{bmatrix}^T, \quad \Gamma = \begin{bmatrix} \gamma_1 & \ldots & \gamma_r \end{bmatrix}^T \tag{7.64d}$$

$$E(t) = \begin{bmatrix} \varepsilon^T(t) & \ldots & \varepsilon^T(t + r - 1) \end{bmatrix}^T \tag{7.64e}$$

Moreover, ℓ is the number, typically equal to r, of input values whose influence on $Y_r(t)$ is to be accounted for. Now, Θ and Γ in (7.63) can be estimated using least squares, giving $\hat{\Theta}_N$ and $\hat{\Gamma}_N$. The k-step ahead predictors are then given by

$$\hat{Y}_r(t) = \hat{\Theta}_N \varphi_s(t) \tag{7.65}$$

For large enough s, this will give a good approximation of the true predictors.

Remark 1: The reason for the term U_t is as follows: The values of $u(t+1), \ldots, u(t+k)$ affect $y(t+k-1)$. If these values can be predicted from past measurements—which is the case if u is not white noise—then the predictions of $y(t+k-1)$ based on past data will account also for the influence of U_t. If we estimate (7.61) directly, this influence will thus be included. However, as demanded by (7.58), the influence of U_t should be ignored in the "formal" k-step ahead predictor we are seeking. This is the reason why this influence is explicitly estimated in (7.62) and then thrown away in the predictor (7.65).

Remark 2: If we seek a state-space realization like (7.55) that does not model the noise properties—*an output error model*—we would just ignore the terms $e(t-j)$ in (7.57)–(7.58). This implies that the predictor in (7.62) would be based on past inputs only, i.e. $s_1 = 0$ in (7.64).

The method thus consists of the following steps:

Basic Subspace Algorithm (7.66)

1. Choose s_1, s_2, r and ℓ and form $\hat{Y}_r(t)$ in (7.65) and $\mathbf{Y}$ as in (7.59).
2. Estimate the rank n of $\mathbf{Y}$ and determine L in (7.60) so that $x(t)$ corresponds to a well-conditioned basis for it.
3. Estimate A, B, C, D and the noise covariance matrices by applying the LS method to the linear regression (7.56).

What we have described now is the *subspace projection* approach to estimating the matrices of the state-space model (7.54), including the basis for the representation and the noise covariance matrices. There are a number of variants of this approach. See among several references, e.g. Van Overschee and DeMoor (1996), Larimore (1983), and Verhaegen (1994).

The approach gives very useful algorithms for model estimation, and is particularly well suited for multivariable systems. The algorithms also allow numerically very reliable implementations, and typically produce estimated models with good quality. If desired, the quality may be improved by using the model as an initial estimate for the prediction error method (7.12). Then the model first needs to be transformed to a suitable parameterization.

The algorithms contain a number of choices and options, like how to choose ℓ, s_i and r, and also how to carry out step number 3. There are also several "tricks" to do step 3 so as to achieve consistent estimates even for finite values of s_i. Accordingly, several variants of this method exist. In Section 10.5 we shall give more algorithmic details around this approach.

7.4 A STATISTICAL FRAMEWORK FOR PARAMETER ESTIMATION AND THE MAXIMUM LIKELIHOOD METHOD

So far we have not appealed to any statistical arguments for the estimation of θ. In fact, our framework of fitting models to data makes sense regardless of a stochastic setting of the data. It is, however, useful and instructive at this point to briefly describe basic aspects of statistical parameter estimation and relate them to our framework.

Estimators and the Principle of Maximum Likelihood

The area of statistical inference, as well as that of system identification and parameter estimation, deals with the problem of extracting information from observations that themselves could be unreliable. The observations are then described as realizations of stochastic variables. Suppose that the observations are represented by the random variable $y^N = (y(1), y(2), \ldots, y(N))$ that takes values in $\mathbf{R}^N$. The probability density function (PDF) of y^N is supposed to be

$$f(\theta; x_1, x_2, \ldots, x_N) = f_y(\theta; x^N) \tag{7.67}$$

That is,

$$P\left(y^N \in A\right) = \int_{x^N \in A} f_y(\theta; x^N)\, dx^N \tag{7.68}$$

In (7.67), θ is a d-dimensional parameter vector that describes properties of the observed variable. These are supposed to be unknown, and the purpose of the observation is in fact to estimate the vector θ using y^N. This is accomplished by an *estimator*,

$$\hat{\theta}(y^N) \tag{7.69}$$

which is a function from $\mathbf{R}^N$ to $\mathbf{R}^d$. If the observed value of y^N is y_*^N, then consequently the resulting estimate is $\hat{\theta}_* = \hat{\theta}(y_*^N)$.

Many such estimator functions are possible. A particular one that maximizes the probability of the observed event is the celebrated maximum likelihood estimator, introduced by Fisher (1912). It can be defined as follows: The joint probability density function for the random vector to be observed is given by (7.67). The probability that the realization (= observation) indeed should take the value y_*^N is thus proportional to

$$f_y(\theta; y_*^N)$$

This is a deterministic function of θ once the numerical value y_*^N is inserted. This function is called the *likelihood function*. It reflects the "likelihood" that the observed event should indeed take place. A reasonable estimator of θ could then be to select it so that the observed event becomes "as likely as possible." That is, we seek

$$\hat{\theta}_{\mathrm{ML}}(y_*^N) = \arg\max_{\theta} f_y(\theta; y_*^N) \tag{7.70}$$

where the maximization is performed for fixed y_*^N. This function is known as the *maximum likelihood estimator* (MLE).

An Example

Let $y(i)$, $i = 1, \ldots, N$, be independent random variables with normal distribution with (unknown) means θ_0 (independent of i) and (known) variances λ_i:

$$y(i) \in N(\theta_0, \lambda_i) \tag{7.71}$$

A common estimator of θ_0 is the sample mean:

$$\hat{\theta}_{\mathrm{SM}}(y^N) = \frac{1}{N}\sum_{i=1}^{N} y(i) \tag{7.72}$$

To calculate the MLE, we start by determining the joint PDF (7.67) for the observations. Since the PDF for $y(i)$ is

$$\frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(x_i - \theta)^2}{2\lambda_i}\right]$$

and the $y(i)$ are independent, we have

$$f_y(\theta; x^N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(x_i - \theta)^2}{2\lambda_i}\right] \tag{7.73}$$

The likelihood function is thus given by $f_y(\theta; y^N)$. Maximizing the likelihood function is the same as maximizing its logarithm. Thus

$$\begin{aligned}\hat{\theta}_{\mathrm{ML}}(y^N) &= \arg\max_{\theta} \log f_y(\theta; y^N) \\ &= \arg\max_{\theta} \left\{-\frac{N}{2}\log 2\pi - \sum_{i=1}^{N}\frac{1}{2}\log\lambda_i - \frac{1}{2}\sum_{i=1}^{N}\frac{(y(i)-\theta)^2}{\lambda_i}\right\}\end{aligned} \tag{7.74}$$

from which we find

$$\hat{\theta}_{\mathrm{ML}}(y^N) = \frac{1}{\sum_{i=1}^{N}(1/\lambda_i)} \sum_{i=1}^{N} \frac{y(i)}{\lambda_i} \tag{7.75}$$

Relationship to the Maximum A Posteriori (MAP) Estimate

The *Bayesian approach* gives a related but conceptually different treatment of the parameter estimation problem. In the Bayesian approach the parameter itself is thought of as a random variable. Based on observations of other random variables that are correlated with the parameter, we may infer information about its value.

Suppose that the properties of the observations can be described in terms of a parameter vector θ. With a Bayesian view we thus consider θ to be a random vector with a certain prior distribution ("prior" means before the observations have been made). The observations y^N are obviously correlated with this θ. After the observations have been obtained, we then ask for the posterior PDF for θ. From this posterior PDF, different estimates of θ can be determined, for example, the value for which the PDF attains its maximum ("the most likely value"). This is known as the *maximum a posteriori (MAP) estimate.*

Suppose that the conditional PDF for y^N, given θ, is

$$f_y(\theta; x^N) = P(y^N = x^N|\theta)$$

and that the prior PDF for θ is

$$g_\theta(z) = P(\theta = z)$$

[Here $P(A|B)$ = the conditional probability of the event A given the event B. We also allowed somewhat informal notation.] Using Bayes's rule (I.10) and with some abuse of notation, we thus find the posterior PDF for θ, i.e., the conditional PDF for θ, given the observations:

$$P(\theta|y^N) = \frac{P(y^N|\theta) \cdot P(\theta)}{P(y^N)} \sim f_y(\theta; y^N) \cdot g_\theta(\theta) \tag{7.76}$$

The posterior PDF as a function of θ is thus proportional to the likelihood function multiplied by the prior PDF. Often the prior PDF has an insignificant influence. Then the MAP estimate

$$\hat{\theta}_{\text{MAP}}(y^N) = \arg\max_\theta \left\{ f_y(\theta; y^N) \cdot g_\theta(\theta) \right\} \tag{7.77}$$

is close to the MLE (7.70).

Cramér-Rao Inequality

The quality of an estimator can be assessed by its mean-square error matrix:

$$P = E\left[\hat{\theta}(y^N) - \theta_0\right]\left[\hat{\theta}(y^N) - \theta_0\right]^T \tag{7.78}$$

Here θ_0 denotes the "true value" of θ, and (7.78) is evaluated under the assumption that the PDF of y^N is $f_y(\theta_0; y^N)$.

We may be interested in selecting estimators that make P small. It is then interesting to note that there is a lower limit to the values of P that can be obtained with various unbiased estimators. This is the so called *Cramér-Rao inequality:*

Let $\hat{\theta}(y^N)$ be an estimator of θ such that $E\hat{\theta}(y^N) = \theta_0$, where E evaluates the mean, assuming that the PDF of y^N is $f_y(\theta_0; y^N)$ (to hold for all values of θ_0), and suppose that y^N may take values in a subset of $\mathbf{R}^N$, whose boundary does not depend on θ. Then

$$E\left[\hat{\theta}(y^N) - \theta_0\right]\left[\hat{\theta}(y^N) - \theta_0\right]^T \geq M^{-1} \tag{7.79}$$

where

$$\begin{aligned} M &= E\left[\frac{d}{d\theta}\log f_y(\theta; y^N)\right]\left[\frac{d}{d\theta}\log f_y(\theta; y^N)\right]^T\Bigg|_{\theta=\theta_0} \\ &= -E\frac{d^2}{d\theta^2}\log f_y(\theta; y^N)\Bigg|_{\theta=\theta_0} \end{aligned} \tag{7.80}$$

Since θ is a d-dimensional vector, $(d/d\theta)\log f_y(\theta; y^N)$ is a d-dimensional column vector and the Hessian $(d^2/d\theta^2)\log f_y(\theta; y^N)$ is a $d \times d$ matrix. This matrix M is known as the *Fisher information matrix*. Notice that the evaluation of M normally requires knowledge of θ_0, so the exact value of M may not be available to the user.

A proof of the Cramér-Rao inequality is given in Appendix 7A.

Asymptotic Properties of the MLE

It is often difficult to exactly calculate properties of an estimator, such as (7.78). Therefore, limiting properties as the sample size (in this case the number N) tends to infinity are calculated instead. Classical such results for the MLE in case of independent observations were obtained by Wald (1949)and Cramér (1946):

Suppose that the random variables $\{y(i)\}$ are independent and identically distributed, so that

$$f_y(\theta; x_1, \ldots, x_N) = \prod_{i=1}^{N} f_{y(i)}(\theta, x_i)$$

Suppose also that the distribution of y^N is given by $f_y(\theta_0; x^N)$ for some value θ_0. Then the random variable $\hat{\theta}_{\mathrm{ML}}(y^N)$ tends to θ_0 with probability 1 as N tends to infinity, and the random variable

$$\sqrt{N}\left[\hat{\theta}_{\mathrm{ML}}(y^N) - \theta_0\right]$$

converges in distribution to the normal distribution with zero mean and covariance matrix given by the Cramér-Rao lower bound [M^{-1} in (7.79) and (7.80)].

In Chapters 8 and 9 we will establish that these results also hold when the ML estimator is applied to dynamical systems. In this sense the MLE is thus the best possible estimator. Let it, however, also be said that the MLE sometimes has been criticized for less good small sample properties and that there are other ways to assess the quality of an estimator than (7.78).

Probabilistic Models of Dynamical Systems

Suppose that the models in the model structure we have chosen in Section 7.1 include both a predictor function and an assumed PDF for the associated prediction errors, as described in Section 5.7:

$$\begin{aligned} \mathcal{M}(\theta): \quad \hat{y}(t|\theta) &= g(t, Z^{t-1}; \theta) \\ \varepsilon(t, \theta) &= y(t) - \hat{y}(t|\theta) \text{ are independent} \\ &\quad \text{and have the PDF } f_e(x, t; \theta) \end{aligned} \tag{7.81}$$

Recall that we term a model like (7.81) that includes a PDF for ε a (complete) *probabilistic model.*

Likelihood Function for Probabilistic Models of Dynamical Systems

We note that, according to the model (7.81), the output is generated by

$$y(t) = g(t, Z^{t-1}; \theta) + \varepsilon(t, \theta) \tag{7.82}$$

where $\varepsilon(t, \theta)$ has the PDF $f_e(x, t; \theta)$. The joint PDF for the observations y^N (given the deterministic sequence u^N) is then given by Lemma 5.1. By replacing the dummy variables x_i by the corresponding observations $y(i)$, we obtain the likelihood function:

$$\begin{aligned} \overline{f}_y(\theta; y^N) &= \prod_{t=1}^{N} f_e\big(y(t) - g(t, Z^{t-1}; \theta), t; \theta\big) \\ &= \prod_{t=1}^{N} f_e(\varepsilon(t, \theta), t; \theta) \end{aligned} \tag{7.83}$$

Maximizing this function is the same as maximizing

$$\frac{1}{N} \log \overline{f}_y(\theta; y^N) = \frac{1}{N} \sum_{t=1}^{N} \log f_e(\varepsilon(t, \theta), t; \theta) \tag{7.84}$$

If we define

$$\ell(\varepsilon, \theta, t) = -\log f_e(\varepsilon, t; \theta) \tag{7.85}$$

we may write

$$\hat{\theta}_{\mathrm{ML}}(y^N) = \arg\min_{\theta} \frac{1}{N} \sum_{t=1}^{N} \ell\,(\varepsilon(t, \theta), \theta, t) \tag{7.86}$$

The maximum likelihood method can thus be seen as a special case of the prediction-error criterion (7.12).

It is worth stressing that (7.85) and (7.86) give the exact maximum likelihood method for the posed problem. It is sometimes pointed out that the exact likelihood function is quite complicated for time-series problems and that one has to resort to approximations of it (e.g., Kashyap and Rao, 1976; Akaike, 1973; Dzhaparidze and Yaglom, 1983). This is true in certain cases. The reason is that it may be difficult to put, say, an ARMA model in the predictor form (7.81) (it will typically require time-varying Kalman predictors). The problem is therefore related to finding the exact predictor and is not a problem with the ML method as such. When we employ time-invariant predictors, we implicitly assume all previous observations to be known [see (3.24)] and typically replace the corresponding initial values by zero or estimate them. Then it is appropriate to interpret the likelihood function as *conditional* w.r.t. these values and to call the method a *conditional ML method* (e.g., Kashyap and Rao, 1976).

Gaussian Special Case

When the prediction errors are assumed to be Gaussian with zero mean values and (t-independent) covariances λ, we have

$$\ell(\varepsilon, \theta, t) = -\log f_e(\varepsilon, t; \theta) = \text{const} + \frac{1}{2}\log\lambda + \frac{1}{2}\frac{\varepsilon^2}{\lambda} \tag{7.87}$$

If λ is known, then (7.87) is equivalent to the quadratic criterion (7.15). If λ is unknown, (7.87) is an example of a parameterized norm criterion (7.16). Depending on the underlying model structure, λ may or may not be parametrized independently of the predictor parameters. See Problem 7E.4 for an illustration of this. Compare also Problem 7E.7.

Fisher Information Matrix and the Cramér-Rao Bound for Dynamical Systems

Having established the log likelihood function in (7.84) for a model structure, we can compute the information matrix (7.80). For simplicity, we then assume that the PDF f_e is known (θ independent) and time invariant. Let $\ell_0(\varepsilon) = -\log f_e(\varepsilon)$. Hence

$$\frac{d}{d\theta}\log \overline{f}_y(\theta; y^N) = \sum_{t=1}^{N} \ell_0'(\varepsilon(t, \theta)) \cdot \psi(t, \theta)$$

where, as in (4.121),

$$\psi(t, \theta) = \frac{d}{d\theta}\hat{y}(t|\theta) = -\frac{d}{d\theta}\varepsilon(t, \theta), \qquad \text{[a } d\text{-dimensional column vector]}$$

Also, ℓ_0' is the derivative of $\ell_0(\varepsilon)$ w.r.t. ε. To find the Fisher information matrix, we now evaluate the expectation of

$$\frac{d}{d\theta}\log \overline{f}_y(\theta; y^N)\left[\frac{d}{d\theta}\log \overline{f}_y(\theta; y^N)\right]^T$$

at θ_0 under the assumption that the true PDF for y^N indeed is $\overline{f}_y(\theta_0; y^N)$. The latter statement means that $\varepsilon(t, \theta_0) = e_0(t)$ will be treated as a sequence of independent random variables with PDF's $f_e(x)$. Call this expectation M_N. Thus

$$\begin{aligned} M_N &= E \sum_{t=1}^{N} \sum_{s=1}^{N} \ell_0'(e_0(t))\, \ell_0'(e_0(s))\, \psi(t, \theta_0) \psi^T(s, \theta_0) \\ &= \sum_{t=1}^{N} E \left[\ell_0'(e_0(t))\right]^2 \cdot E \psi(t, \theta_0) \psi^T(t, \theta_0) \end{aligned}$$

since $e_0(t)$ and $e_0(s)$ are independent for $s \neq t$. We also have $\ell_0'(x) = [\log f_e(x)] = f_e'(x)/f_e(x)$, and

$$\begin{aligned} E \left[\ell_0'(e_0(t))\right]^2 &= \int \frac{\left[f_e'(x)\right]^2}{f_e^2(x)} \cdot f_e(x)\, dx \\ &= \int_{-\infty}^{\infty} \frac{\left[f_e'(x)\right]^2}{f_e(x)} dx \triangleq \frac{1}{\kappa_0} \end{aligned} \tag{7.88}$$

If $e_0(t)$ is Gaussian with variance λ_0, it is easy to verify that $\kappa_0 = \lambda_0$. Hence

$$M_N = \frac{1}{\kappa_0} \cdot \sum_{t=1}^{N} E \psi(t, \theta_0) \psi^T(t, \theta_0) \tag{7.89}$$

Now the Cramér-Rao inequality tells us that for *any unbiased estimator* $\hat{\theta}_N$ of θ (i.e., estimators such that $E\hat{\theta}_N = \theta_0$ regardless of the true value θ_0) we must have

$$\operatorname{Cov} \hat{\theta}_N \geq M_N^{-1} \tag{7.90}$$

Notice that this bound applies for any N and for all parameter estimation methods. We thus have

$$\operatorname{Cov} \hat{\theta}_N \geq \kappa_0 \left[\sum_{t=1}^{N} E \psi(t, \theta_0) \psi^T(t, \theta_0) \right]^{-1} \tag{7.91}$$

$\kappa_0 = \lambda_0$ for Gaussian innovations

Multivariable Gaussian Case (∗)

When the prediction errors are p-dimensional and jointly Gaussian with zero mean and covariance matrices Λ, we obtain from the multivariable Gaussian distribution

$$\ell(\varepsilon, t; \theta) = \text{const} + \tfrac{1}{2} \log \det \Lambda + \tfrac{1}{2} \varepsilon^T \Lambda^{-1} \varepsilon \tag{7.92}$$

Then the negative logarithm of the likelihood function takes the form

$$V_N(\theta, \Lambda, Z^N) = \text{const} + \frac{N}{2} \log \det \Lambda + \frac{1}{2} \sum_{t=1}^{N} \varepsilon^T(t, \theta) \Lambda^{-1} \varepsilon(t, \theta) \tag{7.93}$$

If the $p \times p$ covariance matrix Λ is fully unknown and not parametrized through θ, it is possible to minimize (7.93) analytically with respect to Λ for every fixed θ:

$$\arg\min_{\Lambda} V_N(\theta, \Lambda, Z^N) = \hat{\Lambda}_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t, \theta) \varepsilon^T(t, \theta) \tag{7.94}$$

Then

$$\begin{aligned} \hat{\theta}_N &= \arg\min_{\theta} V_N(\theta, \hat{\Lambda}_N(\theta), Z^N) \\ &= \arg\min_{\theta} \left[\tfrac{1}{2} \log \det \hat{\Lambda}_N(\theta) + \tfrac{1}{2} p \right] \end{aligned} \tag{7.95}$$

(see problem 7D.3) where $p = \dim \varepsilon$. Hence we may in this particular case use the criterion

$$\hat{\theta}_N = \arg\min_{\theta} \det \left[\frac{1}{N} \sum_{t=1}^{N} \varepsilon(t, \theta) \varepsilon^T(t, \theta) \right] \tag{7.96}$$

With this we have actually been led to a criterion of the type (7.29) to (7.30) with $h(A) = \det A$.

Information and Entropy Measures (∗)

In (5.69) and (5.70) we gave a general formulation of a model as an assumed PDF for the observations Z^t:

$$\overline{f}_m(t, Z^t) \tag{7.97}$$

Let $\overline{f}_0(t, Z^t)$ denote the true PDF for the observations. The agreement between two PDF's can be measured in terms of the *Kullback-Leibler information distance* (Kullback and Leibler, 1951):

$$I(\overline{f}_0; \overline{f}_m) = \int \overline{f}_0(t, x^t) \log \frac{\overline{f}_0(t, x^t)}{\overline{f}_m(t, x^t)} \, dx^t \tag{7.98}$$

Here we use x^t as an integration variable for Z^t. This distance is also the *negative entropy* of $\overline{f}_0$ with respect to $\overline{f}_m$:

$$S(\overline{f}_0; \overline{f}_m) = -I(\overline{f}_0; \overline{f}_m) \tag{7.99}$$

An attractive formulation of the identification problem is to *look for a model that maximizes the entropy with respect to the true system* or, alternatively, *minimizes the information distance to the true system.* This formulation has been pursued by Akaike in a number of interesting contributions Akaike (1972, 1974a, 1981).

With a parametrized set of models $\overline{f}_{\mathcal{M}(\theta)}(t, Z^t) = \overline{f}(\theta; t, Z^t)$, we would thus solve

$$\hat{\theta}_N = \arg\min_{\theta} I\left(\overline{f}_0(N, Z^N); \overline{f}(\theta; N, Z^N)\right) \tag{7.100}$$

The information measure can be written

$$\begin{aligned} I(\overline{f}_0; \overline{f}) &= -\int \log\left[\overline{f}(\theta; N, x^N)\right] \cdot \overline{f}_0(N, x^N)\, dx^N \\ &\quad + \int \log\left[\overline{f}_0(N, x^N)\right] \cdot \overline{f}_0(N, x^N)\, dx^N \\ &= -E_0 \log \overline{f}(\theta; N, Z^N) + \theta\text{-independent terms} \end{aligned}$$

where E_0 denotes expectation with respect to the true system.

The problem (7.100) is thus the same as

$$\hat{\theta}_N = \arg\min_{\theta} \left[-E_0 \log \overline{f}(\theta; N, Z^N)\right] \tag{7.101}$$

The problem here is of course that the expectation is not computable since the true PDF is unknown. A simple estimate of the expectation is to replace it by the observation

$$E_0 \log \overline{f}(\theta; N, Z^N) \approx \log \overline{f}(\theta; N, Z^N) \tag{7.102}$$

This gives the log likelihood function for the problem and (7.101) then equals the MLE. The ML approach to identification can consequently also be interpreted as a maximum entropy strategy or a minimum information distance method.

The distance between the resulting model and the true system thus is

$$I\left(\overline{f}_0(N, Z^N); \overline{f}(\hat{\theta}_N; N, Z^N)\right) \tag{7.103}$$

This is a random variable, since $\hat{\theta}_N$ depends on Z^N. As an ultimate criterion of fit, Akaike (1981)suggested the use of the average information distance, or average entropy

$$E_{\hat{\theta}_N} I\left(\overline{f}_0(N, Z^N); \overline{f}(\hat{\theta}_N; N, Z^N)\right) \tag{7.104}$$

This is to be minimized with respect to both the model set and $\hat{\theta}_N$. As an unbiased estimate of the quantity (7.104), he suggested

$$\log \overline{f}(\hat{\theta}_N; N, Z^N) - \dim \theta \tag{7.105}$$

Calculations supporting this estimate will be given in Section 16.4.

The expression (7.105) used in (7.101) gives, with (7.84) and (7.85),

$$\hat{\theta}_{\text{AIC}}(Z^N) = \arg\min_{\theta} \left\{ \frac{1}{N} \sum_{t=1}^{N} \ell(\varepsilon(t,\theta), t, \theta) + \frac{\dim \theta}{N} \right\} \tag{7.106}$$

This is Akaike's information theoretic criterion (AIC). When applied to a given model structure, this estimate does not differ from the MLE in the same structure. The advantage with (7.106) is, however, that the minimization can be performed with respect to different model structures, thus allowing for a general identification theory. See Section 16.4 for a further discussion of this aspect.

An approach that is conceptually related to information measures is Rissanen's minimum description length (MDL) principle. This states that a model should be sought that allows the shortest possible code or description of the observed data. See Rissanen (1978, 1986). Within a given model structure, it gives estimates that coincide with the MLE. See also Section 16.4.

Regularization

Sometimes there is reason to consider the following modified version of the criterion (7.11):

$$W_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \ell(\varepsilon_F(t,\theta)) + \delta|\theta - \theta^{\#}|^2 = V_N(\theta, Z^N) + \delta|\theta - \theta^{\#}|^2 \tag{7.107}$$

It differs from the basic criterion only by adding a cost on the squared distance between θ and $\theta^{\#}$. The latter is a fixed point in the parameter space, and is often taken as the origin, $\theta^{\#} = 0$. The reasons and interpretations for including such a term could be listed as follows:

- If θ contains many parameters, the problem of minimizing V_N may be ill-conditioned, in the sense that the Hessian V_N'' may be an ill-conditioned matrix. Adding the norm penalty will add δI to this matrix, to make it better conditioned. This is the reason why the technique is called *regularization*.
- If the model parameterization contains many parameters (like in the nonlinear black-box models of Section 5.4), it may not be possible to estimate several of them accurately. There are then advantages in pulling them towards a fixed point $\theta^{\#}$. The ones that have the smallest influence on V_N will be affected most by this pulling force. The advantages of this will be brought out more clearly in Section 16.4. We may think of δ as a knob by which we control the effective number of parameters that is used in the minimization. A large value of δ will lock more parameters to the vicinity of $\theta^{\#}$.
- Comparing with the MAP estimate (7.77) we see that this corresponds to minimizing $W_N(\theta, Z^N) = -(1/N)\log\left[\overline{f}_y(\theta, Z^N) \cdot g_\theta(\theta)\right]$ if we take

$$V_N(\theta, Z^N) = -\frac{1}{N} \log \overline{f}_y(\theta; Z^N) \tag{7.108a}$$

$$g_\theta(\theta) = (N\delta/\pi)^{d/2} e^{-N\delta|\theta-\theta^{\#}|^2}, \quad d = \dim\theta \tag{7.108b}$$

that is, we assign a prior probability to the parameters that they are Gaussian distributed with mean θ^* and covariance matrix $\frac{1}{2N\delta}I$. This prior is clearly well in line with the second interpretation.

A Pragmatic Viewpoint

It is good and reassuring to know that general and sound basic principles, such as maximum likelihood, maximum entropy, and minimum information distance, lead to criteria of the kind (7.11). However, in the end we are faced with a sequence of figures that are to be compared with "guesses" produced by the model. It could then always be questioned whether a probabilistic framework and abstract principles are applicable, since we observe only a given sequence of data, and the framework relates to the thought experiment that the data collection can be repeated infinitely many times under "similar" conditions. It is thus an important feature that minimizing (7.11) makes sense, even without a probabilistic framework and without "alibis" provided by abstract principles.

7.5 CORRELATING PREDICTION ERRORS WITH PAST DATA

Ideally, the prediction error $\varepsilon(t, \theta)$ for a "good" model should be independent of past data Z^{t-1}. For one thing, this condition is inherent in a probabilistic model, such as (7.81). Another and more pragmatic way of seeing this condition is that if $\varepsilon(t, \theta)$ is correlated with Z^{t-1} then there was more information available in Z^{t-1} about $y(t)$ than picked up by $\hat{y}(t|\theta)$. The predictor is then not ideal. This leads to the characterization of a good model as one that produces prediction errors that are independent of past data.

A test if $\varepsilon(t, \theta)$ is independent of the whole (and increasing) data set Z^{t-1} would amount to testing whether all nonlinear transformations of $\varepsilon(t, \theta)$ are uncorrelated with all possible functions of Z^{t-1}. This is of course not feasible in practice.

Instead, we may select a certain finite-dimensional vector sequence $\{\zeta(t)\}$ derived from Z^{t-1} and demand a certain transformation of $\{\varepsilon(t, \theta)\}$ to be uncorrelated with this sequence. This would give

$$\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\alpha(\varepsilon(t, \theta)) = 0 \tag{7.109}$$

and the θ-value that satisfies this equation would be the best estimate $\hat{\theta}_N$ based on the observed data. Here $\alpha(\varepsilon)$ is the chosen transformation of ε, and the typical choice would be $\alpha(\varepsilon) = \varepsilon$.

We may carry this idea into a somewhat higher degree of generality. In the first place, we could replace the prediction error with filtered versions as in (7.10). Second, we obviously have considerable freedom in choosing the sequence $\zeta(t)$. It is quite possible that what appears to be the best choice of $\zeta(t)$ may depend on properties of the system. In such a case we would let $\zeta(t)$ depend on θ, and we have the following method:

Choose a linear filter $L(q)$ and let

$$\varepsilon_F(t,\theta) = L(q)\varepsilon(t,\theta) \tag{7.110a}$$

Choose a sequence of correlation vectors

$$\zeta(t,\theta) = \zeta(t, Z^{t-1}, \theta) \tag{7.110b}$$

constructed from past data and, possibly, from θ. Choose a function $\alpha(\varepsilon)$. Then calculate

$$\hat{\theta}_N = \operatorname*{sol}_{\theta \in D_M} \left[f_N(\theta, Z^N) = 0 \right] \tag{7.110c}$$

$$f_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \zeta(t,\theta)\alpha(\varepsilon_F(t,\theta)) \tag{7.110d}$$

Here we used the notation

$$\operatorname{sol}[f(x) = 0] = \text{ the solution(s) to the equation } f(x) = 0$$

Normally, the dimension of ζ would be chosen so that f_N is a d-dimensional vector (which means that ζ is $d \times p$ if the output is a p-vector). Then (7.110) has as many equations as unknowns. In some cases it may be useful to consider an augmented correlation sequence ζ of higher dimension than d so that (7.110) is an overdetermined set of equations, typically without any solution. Then the estimate is taken to be the value that minimizes some quadratic norm of f_N:

$$\hat{\theta}_N = \arg\min_{\theta \in D_M} \left| f_N(\theta, Z^N) \right| \tag{7.111}$$

There are obviously formal links between these correlation approaches and the minimization approach of Section 7.2 (see, e.g., Problem 7D.6).

The procedure (7.110) is a conceptual method that takes different shapes, depending on which model structures it is applied to and on the particular choices of ζ. In the subsequent section we shall discuss the perhaps best known representatives of the family (7.110), the instrumental-variable methods. First, however, we shall discuss the pseudolinear regression models.

Pseudolinear Regressions

We found in Chapter 4 that a number of common prediction models could be written as

$$\hat{y}(t|\theta) = \varphi^T(t,\theta)\theta \tag{7.112}$$

[see (4.21) and (4.45)]. If the data vector $\varphi(t,\theta)$ does not depend on θ, this relationship would be a linear regression. From this the term pseudolinear regression for (7.112) is derived (Solo, 1978). For the model (7.112), the "pseudo-regression vector" $\varphi(t,\theta)$ contains relevant past data, partly reconstructed using the current

model. It is thus reasonable to require from the model that the resulting prediction errors be uncorrelated with $\varphi(t, \theta)$. That is, we choose $\zeta(t, \theta) = \varphi(t, \theta)$ and $\alpha(\varepsilon) = \varepsilon$, in (7.110) and arrive at the estimate

$$\hat{\theta}_N^{\text{PLR}} = \text{sol}\left\{\frac{1}{N}\sum_{t=1}^{N}\varphi(t,\theta)\left[y(t) - \varphi^T(t,\theta)\theta\right] = 0\right\} \tag{7.113}$$

which we term the *PLR estimate*.

Models subject to (7.112) also lend themselves to a number of variants of (7.113), basically corresponding to replacing $\varphi(t, \theta)$ with vectors in which the "reconstructed" (θ-dependent) elements are determined in some other fashion. See Section 10.4.

7.6 INSTRUMENTAL-VARIABLE METHODS

Instrumental Variables

Consider again the linear regression model (7.31):

$$\hat{y}(t|\theta) = \varphi^T(t)\theta \tag{7.114}$$

Recall that this model contains several typical models of linear and nonlinear systems. The least-squares estimate of θ is given by (7.34) and can also be expressed as

$$\hat{\theta}_N^{\text{LS}} = \text{sol}\left\{\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\left[y(t) - \varphi^T(t)\theta\right] = 0\right\} \tag{7.115}$$

An alternative interpretation of the LSE is consequently that it corresponds to (7.110) with $L(q) = 1$ and $\zeta(t, \theta) = \varphi(t)$.

Now suppose that the data actually can be described as in (7.37):

$$y(t) = \varphi^T(t)\theta_0 + v_0(t) \tag{7.116}$$

We then found in Section 7.3 that the LSE $\hat{\theta}_N$ will not tend to θ_0 in typical cases, the reason being correlation between $v_0(t)$ and $\varphi(t)$. Let us therefore try a general correlation vector $\zeta(t)$ in (7.115). Following general terminology in the system identification field, we call such an application of (7.110) to a linear regression an *instrumental-variable* method (IV). The elements of ζ are then called *instruments* or *instrumental variables*. This gives

$$\hat{\theta}_N^{\text{IV}} = \text{sol}\left\{\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\left[y(t) - \varphi^T(t)\theta\right] = 0\right\} \tag{7.117}$$

or

$$\hat{\theta}_N^{\text{IV}} = \left[\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\varphi^T(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\zeta(t)y(t) \tag{7.118}$$

provided the indicated inverse exists. For $\hat{\theta}_N$ to tend to θ_0 for large N, we see from (7.117) that then $(1/N)\sum_{t=1}^{N}\zeta(t)v_0(t)$ should tend to zero. For the method (7.117) to be successfully applicable to the system (7.116), we would thus require the following properties of the instrumental variable $\zeta(t)$ (replacing sample means by expectation):

$$\bar{E}\zeta(t)\varphi^T(t) \text{ be nonsingular} \tag{7.119}$$

$$\bar{E}\zeta(t)v_0(t) = 0 \tag{7.120}$$

In words, we could say that the instruments must be correlated with the regression variables but uncorrelated with the noise. Let us now discuss possible choices of instruments that could be subject to (7.119) and (7.120).

Choices of Instruments

Suppose that (7.114) is an ARX model

$$\begin{aligned} y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) \\ = b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) + v(t) \end{aligned} \tag{7.121}$$

Suppose also that the true description (7.116) corresponds to (7.121) with the coefficients indexed by "zero." A natural idea is to generate the instruments similarly to (7.121) so as to secure (7.119), but at the same time not let them be influenced by $\{v_0(t)\}$. This leads to

$$\begin{aligned} \zeta(t) = K(q)[-x(t-1) \quad -x(t-2)\ldots \\ -x(t-n_a) \quad u(t-1)\ldots u(t-n_b)]^T \end{aligned} \tag{7.122}$$

where K is a linear filter and $x(t)$ is generated from the input through a linear system

$$N(q)x(t) = M(q)u(t) \tag{7.123}$$

Here

$$\begin{aligned} N(q) &= 1 + n_1 q^{-1} + \cdots + n_{n_n} q^{-n_n} \\ M(q) &= m_0 + m_1 q^{-1} + \cdots + m_{n_m} q^{-n_m} \end{aligned} \tag{7.124}$$

Most instruments used in practice are generated in this way. Obviously, $\zeta(t)$ is obtained from past inputs by linear filtering and can be written, conceptually, as

$$\zeta(t) = \zeta(t, u^{t-1}) \tag{7.125}$$

If the input is generated in *open loop* so that it does not depend on the noise $v_0(t)$ in the system, then clearly (7.120) holds. Since both the φ-vector and the ζ-vector are generated from the same input sequence (φ contains in addition effects from v_0), it might be expected that (7.119) should hold "in general." We shall return to this question in Section 8.6.

A simple and appealing choice of instruments is to first apply the LS method to (7.121) and then use the LS-estimated model for N and M in (7.123). The in-

struments are then chosen as in (7.122) with $K(q) = 1$. Systems operating in closed loop and systems without inputs call for other ideas. See Problem 7G.3 for some suggestions.

As outlined in Problem 7D.5, the use of the instrumental vector (7.122) to (7.124) is equivalent to the vector

$$\zeta^*(t) = \frac{K(q)}{N(q)}\left[u(t-1) \quad u(t-2)\dots u(t-n_a-n_b)\right]^T \tag{7.126}$$

The IV estimate $\hat{\theta}_N^{\text{IV}}$ in (7.118) is thus the same for ζ^* as for ζ in (7.122) and does not, for example, depend on the filter M in (7.124).

Model-dependent Instruments (*)

The quality of the estimate $\hat{\theta}_N^{\text{IV}}$ will depend on the choice of $\zeta(t)$. In Section 9.5 we shall derive general expressions for the asymptotic covariance of $\hat{\theta}_N^{\text{IV}}$ and examine them further in Section 15.3. It then turns out that it may be desirable to choose the filter in (7.123) equal to those of the true system: $N(q) = A_0(q)$; $M(q) = B_0(q)$. These are clearly not known, but we may let the instruments depend on the parameters in the obvious way:

$$\begin{aligned}\zeta(t,\theta) &= K(q)\left[-x(t-1,\theta)\dots -x(t-n_a,\theta) \quad u(t-1)\dots u(t-n_b)\right]^T\\ A(q)x(t,\theta) &= B(q)u(t)\end{aligned} \tag{7.127}$$

In general, we could write the generation of $\zeta(t,\theta)$:

$$\zeta(t,\theta) = K_u(q,\theta)u(t) \tag{7.128}$$

where $K_u(q,\theta)$ is a d-dimensional column vector of linear filters.

Including a prefilter (7.110a) and a "shaping" function $\alpha(\cdot)$ for the prediction errors, the IV method could be summarized as follows:

$$\varepsilon_F(t,\theta) = L(q)\left[y(t) - \varphi^T(t)\theta\right] \tag{7.129a}$$

$$\hat{\theta}_N^{\text{IV}} = \operatorname*{sol}_{\theta\in D_{\mathcal{M}}}\left[f_N(\theta, Z^N) = 0\right] \tag{7.129b}$$

where

$$f_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N}\zeta(t,\theta)\alpha(\varepsilon_F(t,\theta)) \tag{7.129c}$$

$$\zeta(t,\theta) = K_u(q,\theta)u(t) \tag{7.129d}$$

Extended IV Methods (*)

So far in this section the dimension of ζ has been equal to dim θ. We may also work with augmented instrumental variable vectors with dimension dim $\zeta > d$. The resulting method, corresponding to (7.110) and (7.111), will be called an *extended IV method* and takes the form

$$\hat{\theta}_N^{\text{EIV}} = \arg\min_{\theta} \left| \frac{1}{N} \sum_{t=1}^{N} \zeta(t, \theta) \alpha(\varepsilon_F(t, \theta)) \right|_Q^2 \tag{7.130}$$

The subscript Q denotes Q-norm:

$$|x|_Q^2 = x^T Q x \tag{7.131}$$

In case ζ does not depend on θ and $\alpha(\varepsilon) = \varepsilon$, (7.130) can be solved explicitly. See Problem 7D.7.

Frequency-domain Interpretation (*)

Quite analogously to (7.20) to (7.25) in the prediction error case, the criterion (7.129) can be expressed in the frequency domain using Parseval's relationship. We then assume that $\alpha(\varepsilon) = \varepsilon$, and that a linear generation of the instruments as in (7.128) is used. This gives

$$f_N(\theta, Z^N) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\hat{\hat{G}}_N(e^{i\omega}) - G(e^{i\omega}, \theta) \right] |U_N(\omega)|^2 \times A(e^{i\omega}, \theta) L(e^{i\omega}) K_u(e^{i\omega}, \theta)\, d\omega \tag{7.132}$$

Here $A(q, \theta)$ is the A-polynomial that corresponds to θ in the model (7.121).

Multivariable Case (*)

Suppose now that the output is p-dimensional and the input m-dimensional. Then the instrument $\zeta(t)$ is a $d \times p$ matrix. A linear generation of $\zeta(t, \theta)$ could still be written as (7.128), with the interpretation that the ith column of $\zeta(t, \theta)$ is given by

$$\zeta^{(i)}(t, \theta) = K_u^{(i)}(q, \theta) u(t) \tag{7.133}$$

where $K_u^{(i)}(q, \theta)$ is a $d \times m$ matrix filter. [$K_u(q, \theta)$ in (7.128) is thus a tensor, a "three-index entity"]. With $\alpha(\varepsilon)$ being a function from $\mathbf{R}^p$ to $\mathbf{R}^p$ and $L(q)$ a $p \times p$ matrix filter, the IV method is still given by (7.129).

7.7 USING FREQUENCY DOMAIN DATA TO FIT LINEAR MODELS (*)

In actual practice, most data are of course collected as samples of the input and output time signals. There are occasions when it is natural and fruitful to consider the Fourier transforms of the inputs and the outputs to be the primary data. It could be, for example, that data are collected by a frequency analyzer, which provides the

transforms to the user, rather than the original time domain data. It could also be that one subjects the measured data to Fourier transformation, before fitting them to models. In some applications, like microwave fields, impedances, etc., the raw measurements are naturally made directly in the frequency domain. This view has been less common in the traditional system identification literature, but has been of great importance in the Mechanical Engineering community, vibrational analysis, and so on. The possible advantages of this will be listed later in this section. The usefulness of such an approach has been made very clear in the work of Schoukens and Pintelon; see in particular the book Schoukens and Pintelon (1991)and the survey Pintelon et.al. (1994).

There is clearly a very close relationship between time domain methods and frequency domain methods for linear models. We saw in (7.25) that the prediction error method for time domain data can (approximately) be interpreted as a fit in the frequency domain. We shall in this section look at some aspects of working directly with data in the frequency domain.

Continuous Time Models

An important advantage to frequency domain data is that it is equally simple to build time continuous models as discrete time/sampled data ones. This means that we can work with models of the kind

$$y(t) = G(p, \theta)u(t) + H(p, \theta)e(t) \tag{7.134}$$

(where p denotes the differentiation operator) analogously to our basic discrete time model (7.3). See also (4.49) and the ensuing discussion. Note the considerable freedom in parameterizing (7.134): from black-box models in terms of numerator and denominator polynomials, or gain, time-delay, and time constant (see (4.50)), to physically parameterized ones like (4.64). In addition to these traditional time-domain parameterizations, one may also parameterize the transfer functions in a way that is more frequency domain oriented. A simple case (see Problem 7G.2) is to let

$$\begin{aligned} G(i\omega, \theta) &= \sum_{k=1}^{d} (g_k^R + i g_k^I) W_\gamma(k, \omega - \omega_k) \\ \theta &= [g_1^R, g_1^I, \ldots, g_d^R, g_d^I] \end{aligned} \tag{7.135}$$

One should typically think of the functions $W_\gamma(k, \omega)$ as bandpass filters, with a width that may be scaled by γ. The parameter g_k would then describe the frequency response around the frequency value ω_k. If the width of the passband increases with frequency we obtain parameterizations linked to wavelet transforms. See, e.g., the insightful discussion by Ninness (1993).

Estimation from Frequency Domain Data

Suppose now that the original data are supposed to be

$$Z^N = \{Y(\omega_k), U(\omega_k), k = 1, \ldots N\} \tag{7.136}$$

where $Y(\omega_k)$ and $U(\omega_k)$ either are the discrete Fourier transforms (2.37) of $y(t)$ and $u(t)$ or are considered as approximations of the Fourier transforms of the underlying continuous signals:

$$Y(\omega) \approx \int_0^\infty y(t)e^{-i\omega t}dt \tag{7.137}$$

Which interpretation is more suitable depends of course on the signal character, sampling interval, and so on.

How to estimate θ in (7.134) or its discrete time counterpart from (7.136)? In view of (7.25) it would be tempting to use

$$\begin{aligned}\hat{\theta}_N &= \arg\min_\theta V(\theta)\\ V(\theta) &= \sum_{k=1}^{N} \left|Y(\omega_k) - G(e^{i\omega_k T}, \theta)U(\omega_k)\right|^2 \cdot \frac{1}{\left|H(e^{i\omega_k T}, \theta)\right|^2}\end{aligned} \tag{7.138}$$

(replacing $e^{i\omega_k T}$ by $i\omega_k$ for the continuous-time model (7.134)). Here T is the sampling interval.

If H in fact does not depend on θ (the case of *fixed* or *known noise model*) experience shows that (7.138) works well. Otherwise the estimate $\hat{\theta}_N$ may not be consistent.

To find a better estimator we turn to the maximum likelihood (ML) method for advice. We give the expressions for the continuous time case; in the case of a discrete time model, just replace $i\omega_k$ by $e^{i\omega_k T}$. We will also be somewhat heuristic with the treatment of white noise.

If the data were generated by

$$y(t) = G(p, \theta)u(t) + H(p, \theta)e(t)$$

the Fourier transforms would be related by

$$Y(\omega) = G(i\omega, \theta)U(\omega) + H(i\omega, \theta)E(\omega) \tag{7.139}$$

To be true, (7.139) should in many cases contain an error term that accounts for finite time effects and the fact that the measured data $Y(\omega_k)$ often are not exact realizations of (7.137). For periodic signals, observed over an integer number of periods, (7.139) may however hold exactly for the input–output relation between u and y.

Now, if $e(t)$ is white noise, its Fourier transform (7.137) will have a complex Normal distribution (see (I.14)):

$$E(\omega) \in N_c(0, \lambda) \tag{7.140}$$

This means that the real and imaginary parts are each normally distributed, with zero means and variances $\lambda/2$. The real and imaginary parts are independent and, moreover, $E(\omega_1)$ and $E(\omega_2)$ are independent for $\omega_1 \neq \omega_2$. (For finite time there will remain some correlation for neighboring frequencies, which we will ignore here.)

This implies that

$$\begin{aligned} &Y(\omega_k) \in N_c(G(i\omega_k,\theta)U(\omega_k), \lambda\,|H(i\omega_k,\theta)|^2) \\ &Y(\omega_k) \text{ and } Y(\omega_\ell) \text{ independent for } \omega_k \neq \omega_\ell \end{aligned} \tag{7.141}$$

according to the model, so that the negative logarithm of the likelihood function becomes

$$\begin{aligned} V_N(\theta) &= N\log\lambda + \sum_{k=1}^{N} 2\log|H(i\omega_k,\theta)| \\ &+ \sum_{k=1}^{N} \frac{1}{\lambda}|Y(\omega_k) - G(i\omega_k,\theta)U(\omega_k)|^2 \cdot \frac{1}{|H(i\omega_k,\theta)|^2} \end{aligned} \tag{7.142}$$

The Maximum Likelihood estimate is

$$\hat{\theta}_N = \arg\min_{\theta} V_N(\theta) \tag{7.143}$$

Remark: This is the ML criterion under the assumption (7.141). We noted above that the data might not be exactly subject to this condition, due to finite time effects when forming the Fourier transforms. It still makes sense to use the criterion (7.143), though.

If we perform analytical minimization of (7.143) w.r.t. λ, we obtain

$$\hat{\theta}_N = \arg\min_{\theta}\left[N \cdot \log W_N(\theta) + 2\sum_{k=1}^{N}\log|H(i\omega_k,\theta)|\right] \tag{7.144}$$

$$W_N(\theta) = \frac{1}{N}\sum_{k=1}^{N}|Y(\omega_k) - G(i\omega_k,\theta)U(\omega_k)|^2 \cdot \frac{1}{|H(i\omega_k,\theta)|^2} \tag{7.145}$$

$$\hat{\lambda}_N = W_N(\hat{\theta}_N) \tag{7.146}$$

Compared to (7.138) we thus have an extra term

$$\sum_{k=1}^{N}\log|H(i\omega_k,\theta)|^2 \tag{7.147}$$

If the noise model is given and fixed, H does not depend on θ, and the term (7.147) does not affect the estimate. This case of fixed noise models is very common in applications with frequency domain data (see Schoukens and Pintelon, 1991). One reason is that for a periodic input, we can obtain reasonable estimates of H in a preprocessing step. See Schoukens et.al. (1997)and (7.154) below.

We may also note that for any monic, stable, and inversely stable transfer function $H(q, \theta)$ we have

$$\int_{-\pi}^{\pi} \log \left| H(e^{i\omega}, \theta) \right|^2 d\omega \equiv 0 \tag{7.148}$$

The expression will also hold if the integral is replaced by summation over the frequencies $2\pi k/N$, $k = 1, \ldots, N$. This is the reason why (7.147) is missing from time domain criteria, like (7.25), which correspond to equally spaced frequencies ω_k.

Note that $W_N(\theta)$ in (7.144) can be rewritten as

$$W_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{\hat{G}}(i\omega_k) - G(i\omega_k, \theta) \right|^2 \cdot \frac{|U(\omega_k)|^2}{|H(i\omega_k, \theta)|^2} \tag{7.149}$$

in formal agreement with (7.25). Here $\hat{\hat{G}}$ is the empirical transfer function estimate, ETFE, defined in (6.24).

Some Variants of the Criterion

Weighted Nonlinear Least Squares Criterion. Given an estimate $\hat{\hat{G}}$ of the frequency function (the ETFE or anything else), it is natural to fit a parametric model to it by a (non-linear) least squares criterion

$$W_N^{\mathrm{NLS}}(\theta) = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{\hat{G}}(i\omega_k) - G(i\omega_k, \theta) \right|^2 W_k \tag{7.150}$$

with some weighting function W_k. We see that this corresponds to the ML criterion with $W_k = |U(\omega_k)|^2 / |H(i\omega_k, \theta)|^2$. In other words, (7.150) can be interpreted as the ML criterion with a fixed noise model

$$|H(i\omega_k)|^2 = \frac{|U(\omega_k)|^2}{W_k} \tag{7.151}$$

The numerical minimization of this criterion is typically carried out using a damped Gauss-Newton method, like for most of the other criteria discussed in this book. See Section 10.2.

A Linear Method. If we use an ARX-parameterization of the model (see (4.9))

$$G(p, \theta) = \frac{B(p)}{A(p)}, \qquad H(p, \theta) = \frac{1}{A(p)}$$

then the criterion (7.149) takes the form

$$W_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} \left| A(i\omega_k) \hat{\hat{G}}(i\omega_k) - B(i\omega_k) \right|^2 |U(\omega_k)|^2 \tag{7.152}$$

This is a quadratic criterion in the coefficients of the polynomials A and B. It can therefore be minimized explicitly by the least squares solution.

A number of other variants can also be defined. We can, for example, define a frequency domain IV-method in analogy with (7.132). See Pintelon et.al. (1994)for a more complete survey.

Some Practical Features with Estimation from Frequency Domain Data

There are several distinct features with the direct frequency domain approach that could be quite useful. We shall list a few:

- **Prefiltering** is known as quite useful in the time-domain approach. See Section 14.4. For frequency domain data it becomes very simple: It just corresponds to assigning different weights to different frequencies in the weighted criterion (7.150). This, in turn, is the same as invoking a special noise model (7.151).

  Normally, it does not quite make sense to combine prefiltering with estimating a noise model, since a parameter-dependent weighting as in

$$W_N^{\mathrm{NLS}}(\theta) = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{\hat{G}}(i\omega_k) - G(i\omega_k,\theta)\right|^2 W_k\frac{|U(\omega_k)|^2}{|H(i\omega_k,\theta)|^2}$$

  may undo any applied weighting from W.
- **Condensing Large Data Sets.** When dealing with systems with a fairly wide spread of time constants, large data sets have to be collected in the time domain. When converted to the frequency domain they can easily be condensed, so that, for example, logarithmically spaced frequencies are obtained. At higher frequencies one would thus decimate the data, which involves averaging over neighboring frequencies. Then the noise level (λ_k) is reduced accordingly.
- **Combining Experiments.** Nothing in the approach of (7.141)–(7.143) says that the frequency response data at different frequencies have to come from the same experiment, or even that the frequencies involved ($\omega_k, k = 1, \ldots, N$) all have to be different. It is thus very easy to combine data from different experiments.
- **Periodic Inputs.** The main drawback with the frequency domain approach is that the underlying frequency domain model (7.139) is strictly correct only for a periodic input and assuming all transients have died out. On the other hand, typical use of the time domain method assumes inputs and outputs prior to time $t = 0$ to be zero. Whichever assumption about past behavior is closer to the truth should thus affect the choice of approach. Note, though, that both the time-domain and the frequency-domain methods allow the possibility to estimate a finite number of parameters that pick up these transients, and thus give correct handling of these effects. See also Section 13.3.
- **Non-Parametric Noise Estimates from Periodic Inputs.** We have for the true system $y(t) = y_u(t) + v(t)$, where $y_u(t) = G_0(q)u(t)$. If $u(t)$ is periodic with period M, so will $y_u(t)$ be, after a transient. By averaging the output over K periods,

$$\bar{y}(t) = \frac{1}{K}\sum_{k=0}^{K-1} y(t + kM), \quad t = 1, \ldots, M \tag{7.153}$$

we can thus get a better estimate of $y_u(t)$, and also estimate the noise sequence as

$$\hat{v}(t) = y(t) - \bar{y}(t) \tag{7.154}$$

where the definition of $\bar{y}(t)$ has been extended to $1 \le t \le N$ by periodic continuation. Estimating the spectrum of $\hat{v}(t)$ with any (non-parametric) method gives a noise spectral model $|H(i\omega)|^2$ that can be used in (7.144).

- **Band-Limited Signals.** If the actual input signals are band-limited (like no power above the Nyquist frequency), the continuous time Fourier transform (7.137) can be well computed from sampled data. It is then possible to directly build continuous-time models without any extra work. Notice also that frequency contents above the Nyquist frequency can be eliminated from both input and output signals by anti-alias filtering (see Section 13.7) before sampling. Such filtering will not distort the input-output relationship, provided the input and output are subjected to exactly the same filters.
- **Continuous-Time Models.** The comment above shows that direct continuous-time system identification from "continuous-time data" can be dealt with in a rather straightforward fashion. Otherwise, continuous time data with continuous-time white noise descriptions are delicate mathematical objects.
- **Trade-off Noise/Frequency Resolution.** The approach also allows for a more direct and frequency dependent trade-off between frequency resolution and noise levels. That will be done as the original Fourier transform data are decimated to the selected range of frequencies ω_k, $k = 1, \ldots, N$.

7.8 SUMMARY

There are several ways to fit models in a given set to observed data. In this chapter we have pointed out two general procedures. Both deal with the sequence of prediction errors $\{\varepsilon(t, \theta)\}$ computed from the respective models using the observed data, and both could be said to aim at making this sequence "small."

The *prediction-error identification approach* (PEM) was defined by (7.10) to (7.12):

$$\hat{\theta}_N = \arg\min_{\theta \in D_M} V_N(\theta, Z^N)$$

$$V_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \ell(\varepsilon(t, \theta), \theta, t) \tag{7.155}$$

It contains well-known procedures, such as the least-squares (LS) method and the maximum-likelihood (ML) method and is at the same time closely related to Bayesian maximum a posteriori (MAP) estimation and Akaike's information criterion (AIC).

The *subspace approach to identifying state-space models* was defined by (7.66). It consists of three steps: (1) estimating the k-step ahead predictors using an LS-algorithm, and (2) selecting the state vector from these, and finally (3) estimating the state-space matrices using these states and the LS-method.

The *correlation approach* was defined by (7.110):

$$\varepsilon_F(t,\theta) = L(q)\varepsilon(t,\theta)$$

$$\hat{\theta}_N = \operatorname*{sol}_{\theta\in D_{\mathcal{M}}} \left[f_N(\theta, Z^N) = 0\right]$$

$$f_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \zeta(t,\theta)\alpha(\varepsilon_F(t,\theta)) \tag{7.156}$$

It contains the instrumental-variable (IV) technique, as well as several methods for rational transfer function models.

System identification has often been described as an area crowded with seemingly unrelated *ad hoc* methods and tricks. The list of names of available and suggested methods is no doubt a very long one. It is our purpose, however, with this chapter, as well as Chapters 8 to 11, to point out that the number of underlying basic ideas is really quite small, and that it indeed is quite possible to orient oneself in the area of system identification with these basic ideas as a starting point.

It might be added that for systems operating in closed loop some special identification techniques have been devised. We shall review these methods in Section 13.5, in connection with a discussion of the closed loop experiment situation. The bottom line is that a direct application of the prediction error methods of this chapter should be the prime choice, also for closed loop data.

7.9 BIBLIOGRAPHY

The parameter estimation methods described in this chapter all go back to basic methods in statistics. For general texts, we refer to Cramér (1946), Rao (1973), and Lindgren (1976).

Section 7.2: The term "prediction-error-methods" was perhaps first coined in Ljung (1974), but it had long been realized that the common methods of system identification had aimed at making the prediction error small. From an operational point of view, the criterion (7.155) can be viewed as a nonlinear regression method. See, for example, Jennrich (1969)and Hannan (1971b). Various norms have been discussed (see also Section 15.2). The ℓ_∞-norm, which is related to *unknown-but-bounded disturbances*, and *set membership identification*, is discussed in early contributions by Schweppe (1973)and Fogel and Huang (1982)(cf. Problem 7G.7). This approach has then been discussed in numerous papers, see the surveys in Milanese and Vicino (1991), Deller (1990), and Walter and Piet-Lahanier (1990). See also the bibliographies of Chapters 14 and 16 for contributions to this approach, that relate to identification for control and model validation.

The frequency-domain expressions for the prediction-error criteria go back to Whittle (1951), who dealt with the input-free case. Among many references, we may mention Hannan (1970), Chapter VI, for a detailed study (still with no input). Related formulas are also given by Solo (1978), Ljung and Glover (1981), and Wahlberg and Ljung (1986). An alternative approach based on interpolation of a given frequency function has become known as $\mathcal{H}_\infty$-identification. See e.g., Gu and Khargonekar (1992)and Helmicki, Jacobson, and Nett (1991).

Section 7.3: The statistical roots of the least-squares method are examined in Appendix II. The application to times series has its origin in the work of Yule (1927)and Walker (1931), with a first asymptotic analysis by Mann and Wald (1943). The application to dynamic systems with an input was made independently by several authors, with an early comprehensive description and analysis by Åström (1968), to some extent reprinted in Åström and Eykhoff (1971). A good account of different variants of the LS method is given in Hsia (1977). *The Subspace methods* really originate from classical realization theory as formulated in Ho and Kalman (1966)and Kung (1978). A basic treatment of the modern techniques is given in the book Van Overschee and DeMoor (1996). More references will be listed in Chapter 10 in connection with the algorithmic details of the approach.

Section 7.4: Whittle pioneered maximum likelihood methods for AR and ARMA models using frequency domain formulations, see e.g., Whittle (1951). The principle of maximum likelihood was then applied to dynamical systems by Åström and Bohlin (1965)(ARMAX model structures) and Box and Jenkins (1970)[model structure (4.31)]. Since then a long list of articles has dealt with this approach. Åström (1980)may be singled out for a survey.

Frequency-domain variants or approximations of the likelihood function have been extensively used by Whittle (1951), Hannan (1970), and others. The Bayesian MAP approach is comprehensively treated in Peterka (1981a), and Peterka (1981b). The calculations leading to (7.96) were first given by Eaton (1967)and Akaike (1973). Entropy and information theoretic criteria have been discussed extensively by Akaike and Rissanen. We may single out Akaike (1974a), Akaike (1981), Rissanen (1985), and Rissanen (1986)as recommended reading. A general reference on entropy and statistics is Kullback (1959). Regularization is really a general technique to solve ill-posed problems, Tikhonov and Arsenin (1977). Its applications to estimation is discussed in Wahba (1987), Wahba (1990), and Sjöberg, McKelvey, and Ljung (1993). Regularization is of particular importance for nonlinear black box models, where many parameters often are used. See, Sjöberg et.al. (1995)and Girosi, Jones, and Poggio (1995). The connection to Bayesian approaches in the latter context are pursued in Williams (1995), while Johansen (1996)describes how regularization can be used to incorporate prior knowledge in nonlinear models. See also (16.36) in Chapter 16.

Sections 7.5–7.6: The way to describe the "correlation approach" as presented here is novel, although the different methods are well known. The IV method was introduced into statistics and econometrics by Reiersøl (1941)and has been applied to many parameter estimation problems in statistics (see, e.g., Kendall and Stuart, 1961). Applications to dynamic systems in the control field have been pioneered by Wong and Polak (1967), Young (1965), and Mayne (1967). For applications to ARMA models see Stoica, Friedlander and Söderström (1986). A historical background is given by Young (1976). For comprehensive treatments, see Söderström and Stoica (1983)and Young (1984).

Section 7.8: An early reference for frequency-domain criteria of the type (7.152) is Levi (1959). A comprehensive treatment with many references is the book Schoukens and Pintelon (1991). See also the survey paper Pintelon et.al. (1994). There is a MATLAB Toolbox devoted to frequency domain identification methods, Kollàr (1994).

7.10 PROBLEMS

7G.1 *Input error and output error methods:* Consider a model structure

$$y(t) = G(q, \theta)u(t)$$

without a specified noise model. In the survey of Åström and Eykhoff (1971) identification methods that minimize "the output error"

$$\hat{\theta}_N = \arg\min \sum_{t=1}^{N} [y(t) - G(q, \theta)u(t)]^2$$

and the "input error"

$$\hat{\theta}_N = \arg\min \sum_{t=1}^{N} \left[u(t) - G^{-1}(q, \theta)y(t)\right]^2$$

are listed. Show that these methods are prediction error methods corresponding to particular choices of noise models $H(q, \theta)$.

7G.2 *Spectral analysis as a prediction error method:* Consider the model structure

$$G(e^{i\omega}, \theta) = \sum_{k=1}^{N} \left(g_k^R + ig_k^I\right) W_\gamma(\omega - \omega_k)$$

$$\theta = \begin{bmatrix} g_1^R & g_1^I, \ldots, g_n^R & g_n^I \end{bmatrix}^T$$

and let $H(e^{i\omega}, \eta)$ be an arbitrary noise model parametrization. Let $\hat{\theta}_N$ be the prediction-error estimate obtained by minimization of (7.23) and (7.25):

$$\hat{\theta}_N = \arg\min_{\theta, \eta} \int_{-\pi}^{\pi} \frac{\left|\hat{\hat{G}}_N(e^{i\omega}) - G(e^{i\omega}, \theta)\right|^2 |U_N(\omega)|^2}{\left|H(e^{i\omega}, \eta)\right|^2} d\omega$$

(a) Consider the special case $H(e^{i\omega}, \eta) \equiv 1$ and

$$W_\gamma(\omega) = \begin{cases} 1, & |\omega| \le \dfrac{\pi}{2n} \\ 0, & |\omega| > \dfrac{\pi}{2n} \end{cases}$$

$$\omega_k = \frac{(k-1)\pi}{n}$$

Show that $\hat{G}(e^{i\omega_k}, \hat{\theta}_N)$ is then given by (6.46).

(b) Assume, in the general case, that

$$H(e^{i\omega}, \eta) \cdot W_\gamma(\omega - \omega_k) \approx H(e^{i\omega_k}, \eta) \cdot W_\gamma(\omega - \omega_k)$$

$$G(e^{i\omega}, \theta) \cdot W_\gamma(\omega - \omega_k) \approx G(e^{i\omega_k}, \theta) \cdot W_\gamma(\omega - \omega_k)$$

Show that (6.46) then holds approximately.

7G.3 *Instruments for closed-loop systems:* Consider a system

$$A_0(q)y(t) = B_0(q)u(t) + v_0(t)$$

under the output feedback

$$u(t) = F_1(q)r(t) - F_2(q)y(t)$$

(a) Let $x(t)$ and $\zeta(t)$ be given by

$$N(q)x(t) = M(q)r(t)$$

$$\zeta(t) = K(q)\left[-x(t-1)\dots -x(t-n_a) \quad r(t-1)\dots r(t-n_b)\right]^T$$

Show that (7.120) holds for these instruments, and verify that (7.119) holds for a simple first-order special case.

(b) Suppose that $v_0(t)$ is known to be an MA process of order s. Introduce the instruments

$$\zeta(t) = [-y(t-1-s)\dots -y(t-n_a-s) \quad u(t-1-s)\dots u(t-n_b-s)]^T$$

Show the same results as under part (a). See also Söderström, Stoica and Trulsson (1987).

7G.4 Suppose $Y_N = [y(1), \dots, y(N)]^T$ is a Gaussian N-dimensional random vector with zero mean and covariance matrix $R_N(\theta)$. Let

$$R_N(\theta) = L_N(\theta)\Lambda_N(\theta)L_N^T(\theta)$$

where $L_N(\theta)$ is lower triangular with 1's along the diagonal and $\Lambda_N(\theta)$ a diagonal matrix with $\lambda_\theta(t)$ as the t, t element. Let

$$E_N(\theta) = L_N^{-1}(\theta)Y_N$$

$$E_N(\theta) = [\varepsilon(1,\theta), \dots, \varepsilon(N,\theta)]^T$$

Show that, if θ is a parameter to be estimated, then the negative log likelihood function when Y_N is observed is

$$\frac{N}{2}\log 2\pi + \frac{1}{2}\log\det R_N(\theta) + \frac{1}{2}Y_N^T R_N^{-1}(\theta)Y_N$$

Show also that this can be rewritten as

$$\frac{N}{2}\log 2\pi + \frac{1}{2}\sum_{t=1}^{N}\log\lambda_\theta(t) + \frac{1}{2}\sum_{t=1}^{N}\frac{\varepsilon^2(t,\theta)}{\lambda_\theta(t)}$$

where $\varepsilon(t,\theta)$ are independent, normal random variables with variances $\lambda_\theta(t)$. How does this relate to our calculations (7.81) to (7.87)?

7G.5 Let the two random vectors X and Y be jointly Gaussian with

$$EX = m_X; \qquad EY = m_Y$$
$$E(X - m_X)(X - m_X)^T = P_X \qquad E(Y - m_Y)(Y - m_Y)^T = P_Y$$
$$E(X - m_X)(Y - m_Y)^T = P_{XY}$$

Show that the conditional distribution of X given Y is

$$(X|Y) \in N\left(m_X + P_{XY}P_Y^{-1}(Y - m_Y),\ P_X - P_{XY}P_Y^{-1}P_{XY}^T\right)$$

7G.6 Consider the model structure

$$\begin{aligned} X &= F(\theta)W \\ Y &= H(\theta)X + E \end{aligned} \tag{7.157}$$

where W and E are two independent, Gaussian random vectors with zero mean values and unit covariance matrices. Note that state-space models like (4.84), without input, can be written in this form by forming $X^T = [x^T(1) \quad x^T(2) \ldots x^T(N)]$ and $Y^T = [y(1) \quad y(2) \ldots y(N)]$. Let

$$R(\theta) = I + H(\theta)F(\theta)F^T(\theta)H^T(\theta)$$

Show the following:

(a) The negative log likelihood function for θ, (ignoring θ-independent terms) when Y is observed is

$$V(\theta) = -\log p(Y|\theta) = \tfrac{1}{2}Y^T R^{-1}(\theta)Y + \tfrac{1}{2}\log\det R(\theta)$$

Let

$$\hat{\theta}_{\mathrm{ML}} = \arg\min_{\theta} V(\theta)$$

(cf. Problem 7G.4).

(b) Let the conditional expectation of X, given Y and θ be $\hat{X}^s(\theta)$. Show that

$$E(X|Y, \theta) = \hat{X}^s(\theta) = \left[F(\theta)F^T(\theta)H^T(\theta)\right]R^{-1}(\theta)Y \tag{7.158}$$

and that

$$-\log p(X|\theta, Y) = \tfrac{1}{2}\left(X - \hat{X}^s(\theta)\right)^T S^{-1}(\theta)\left(X - \hat{X}^s(\theta)\right) + \tfrac{1}{2}\log\det S(\theta)$$
$$S(\theta) = F(\theta)F^T(\theta) - F(\theta)F^T(\theta)H^T(\theta)R^{-1}(\theta)H(\theta)F(\theta)F^T(\theta) \tag{7.159}$$

(cf. Problem 7G.5) [$\hat{X}^s(\theta)$ gives the *smoothed* state estimate for the underlying state space model, see Anderson and Moore (1979)].

(c) Assume that the prior distribution of θ is flat. ($p(\theta) \approx$ independent of θ). Then show that the joint MAP estimate (7.77) of θ and X given Y,

$$(\hat{\theta}^s_{\text{MAP}}, \hat{X}^s_{\text{MAP}}) = \arg\max_{\theta, X} p(\theta, X|Y)$$

is given by

$$\arg\min_{X,\theta} [-\log p(Y, X|\theta)]$$

where

$$-\log p(Y, X|\theta) = \tfrac{1}{2}|Y - H(\theta)X|^2 + \tfrac{1}{2}|F^{-1}(\theta)X|^2 + \log\det F(\theta) \tag{7.160}$$

(d) Show that the value of X that minimizes (7.160) for fixed Y and θ is $\hat{X}^s(\theta)$, defined by (7.158). Hence

$$\hat{\theta}^s_{\text{MAP}} = \arg\min_{\theta} \left\{ \tfrac{1}{2}\left|Y - H(\theta)\hat{X}^s(\theta)\right|^2 + \tfrac{1}{2}\left|F^{-1}(\theta)\hat{X}^s(\theta)\right|^2 + \log\det F(\theta) \right\}$$

$$\hat{X}^s_{\text{MAP}} = \hat{X}^s(\hat{\theta}^s_{\text{MAP}})$$

(e) Establish that

$$-\log p(Y|\theta) = -\log p(Y, X|\theta) + \log p(X|\theta, Y) \tag{7.161}$$

(f) Establish that

$$-\log p(Y|\theta) = \left[\tfrac{1}{2}\left|Y - H(\theta)\hat{X}^s(\theta)\right|^2 + \tfrac{1}{2}\left|F^{-1}(\theta)\hat{X}^s(\theta)\right|^2 + \tfrac{1}{2}\log\det R(\theta) \right] \tag{7.162}$$

[*Hint:* Use the matrix identity (cf. (7.159))

$$S(\theta) = \left[F^{-T}(\theta)F^{-1}(\theta) + H^T(\theta)H(\theta)\right]^{-1}$$

and the determinant identity

$$\det(I_r + AB) = \det(I_s + BA)$$

for A and B being $r \times s$ and $s \times r$ matrices and I_r the $r \times r$ identity matrix.]

(g) Conclude that $\hat{\theta}_{\text{ML}} \neq \hat{\theta}^s_{\text{MAP}}$ in general.

Remark: The problem illustrates the relationships among various expressions for the likelihood function, the smoothing problem, and MAP-estimates. "Log likelihood functions" of the kind (7.160) have been discussed, e.g., in Sage and Melsa (1971)and Schweppe (1973), Section 14.3.2.

7G.7 Consider the linear regression structure

$$y(t) = \varphi^T(t)\theta + v(t)$$

Based on the theory of optimal algorithms for operator approximation, (Traub and Wozniakowski, 1980), Milanese and Tempo (1985), and Milanese, Tempo, and Vicino (1986)have suggested the following estimate:

For given δ, y_N and $\{\varphi(t)\}_1^N$, define the set

$$A_\delta = \{\theta \mid |y(t) - \varphi^T(t)\theta| < \delta \quad \text{all } t = 1, \ldots, N\}$$

Assuming A_δ to be bounded and non-empty define its "center" $\theta_c(A_\delta)$ as follows: The ith component is

$$[\theta_c(A_\delta)]^{(i)} = \tfrac{1}{2}\left[\sup_{\theta \in A_\delta} \theta^{(i)} + \inf_{\theta \in A_\delta} \theta^{(i)}\right]$$

(superscript (i) denoting i : th component). The estimate $\hat{\theta}_N^\delta$ is then taken as $\theta_c(A_\delta)$.

(a) Suppose that dim $\theta = 1$. Prove that $\hat{\theta}_N^\delta$ is independent of δ, as long as A_δ is nonempty and bounded.

(b) When dim $\theta > 1$, $\hat{\theta}_N^\delta$ may in general depend on δ. Suppose that as δ decreases to a value δ^*, A_δ reduces to a singleton

$$A_{\delta^*} = \{\theta^\dagger\}$$

Then clearly $\hat{\theta}_N^{\delta^*} = \theta^\dagger$. Show that

$$\hat{\theta}_N^{\delta^*} = \arg\min_\theta \max_t |y(t) - \varphi^T(t)\theta|$$

This "optimal estimate" thus corresponds to the prediction error estimate (7.12) with the ℓ_∞-norm

$$\ell_\infty\left(\varepsilon^N(\cdot, \theta)\right) = \max_t |\varepsilon(t, \theta)|$$

This in turn can be seen as the limit as $p \to \infty$ of the criterion functions

$$\ell(\varepsilon) = |\varepsilon|^p$$

in (7.11).

7E.1 *Estimating the AR Part of an ARMA model:* Consider the ARMA model

$$A(q)y(t) = C(q)e(t)$$

with orders n_a and n_c, respectively. A method to estimate the AR part has been given as follows. Let

$$\hat{R}_y^N(\tau) = \frac{1}{N}\sum_{t=\tau}^{N} y(t)y(t-\tau)$$

Then solve for $\hat{a}_i^N$ from

$$\hat{R}_y^N(\tau) + a_1\hat{R}_y^N(\tau - 1) + \cdots + a_{n_a}\hat{R}_y^N(\tau - n_a) = 0$$

$$\tau = n_c + 1, n_c + 2, \ldots, n_c + n_a$$

Show that this (essentially) is an application of the IV method using specific instruments. Which ones? (See Cadzow, 1980, and Stoica, Söderström, and Friedlander, 1985.)

7E.2 *Sinusoids in noise:* Consider a sinusoid measured in white Gaussian noise:

$$y(t) = \alpha e^{i\omega t} + e(t)$$

For simplicity we use complex algebra. The constant α is thus complex-valued. The amplitude, phase and frequency are unknown: $\theta = (\alpha, \omega)$. The predictor thus is

$$\hat{y}(t|\theta) = \alpha e^{i\omega t}$$

If $e(t)$ has variance 1 (real and imaginary parts independent), the likelihood function gives the prediction-error criterion:

$$V_N(\theta, Z^N) = \frac{1}{2}\sum_{t=1}^{N} \left|y(t) - \hat{y}(t|\theta)\right|^2$$

Show that the MLE

$$\hat{\theta}_N = \begin{bmatrix} \hat{\alpha}_N \\ \hat{\omega}_N \end{bmatrix} = \arg\min_{\theta} V_N(\theta, Z^N)$$

obeys

$$\hat{\omega}_N = \arg\max_{\omega} |Y_N(\omega)|^2$$

where $Y_N(\omega)$ is the Fourier transform (2.37) of $y(t)$.

7E.3 *Error-in-variables models:* Econometric models often include disturbances both on inputs and outputs (compare our comment in Section 2.1 on Figure 2.2). Consider the model in Figure 7.1. The true inputs and outputs are thus s and x, while we measure u and y. In a first-order case, we have

$$\begin{aligned} x(t) + ax(t-1) &= bs(t-1) \\ y(t) &= x(t) + e(t) \\ u(t) &= s(t) + w(t) \end{aligned}$$

Suppose that w and e are independent white noises with unknown variances. Discuss how a, b, and these variances can be estimated using measurements of y and u.

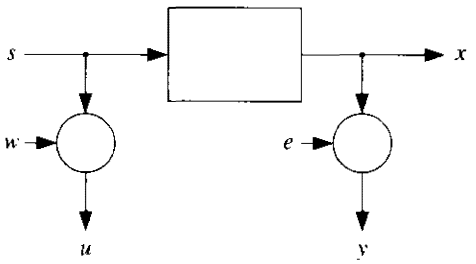

Figure 7.1 An error-in-variables model.

[*Remark:* With the assumption that the color of the noises are known, the problem is relatively simple. Without this assumption the problem is more difficult. See Kalman (1991), Anderson (1985), Söderström (1981), McKelvey (1996), and Stoica et.al. (1997)].

7E.4 Consider a probabilistic model, implicitly given in the state-space form

$$\begin{aligned} x(t+1) &= ax(t) + w(t) \\ y(t) &= x(t) + v(t) \end{aligned} \tag{7.163a}$$

where $\{w(t)\}$ and $\{v(t)\}$ are assumed to be independent, white Gaussian noises, with variances

$$\begin{aligned} Ew^2(t) &= r_1 \\ Ev^2(t) &= 1 \quad \text{(assumed known)} \end{aligned} \tag{7.163b}$$

Let the parameter vector be

$$\theta = \begin{bmatrix} a \\ r_1 \end{bmatrix} \tag{7.163c}$$

Assume initial conditions for $x(0)$ (mean and variance) such that the prediction $\hat{y}(t|\theta)$ becomes a stationary process for each θ (i.e., so that the steady-state Kalman filter can be used). Determine the log-likelihood function for this problem. Compare with the log-likelihood function for a directly parametrized innovations representation model (4.91).

7E.5 Consider the nonlinear model structure of Problem 5E.1. Discuss how the LS, ML, IV, and PLR methods can be applied to this structure. (Reference: Fnaiech and Ljung, 1986).

7E.6 Consider the model structure

$$y(t) = \varphi^T(t)\theta + v(t)$$

where the regression vector $\varphi(t)$ can only be measured with noise:

$$\eta(t) = \varphi(t) + w(t)$$

The noises $\{w(t)\}$ and $\{v(t)\}$ may be nonwhite and mutually correlated. Suppose a vector $\zeta(t)$ is known that is uncorrelated with $\{v(t)\}$ and $\{w(t)\}$ but correlated with $\varphi(t)$. Suggest how to estimate θ from $y(t)$, $\eta(t)$, and $\zeta(t)$, $t = 1, \ldots, N$.

7E.7 Suppose in (7.86) and (7.87) that λ does not depend on θ. Determine $\hat{\lambda}_N$.

7E.8 Consider the model structure

$$\hat{y}(t|\theta) = -ay(t-1) + bu(t-1)$$

and assume that the true system is given by

$$y(t) - 0.9y(t-1) = u(t-1) + e_0(t)$$

where $\{e_0(t)\}$ is white noise of unit variance. Determine the Cramér-Rao bound for the estimation of a and b. How does it depend on the properties of u?

7E.9 Suppose that $u(t)$ is periodic with period M, and that all transients have died out. We collect data over K periods: $[y(t), u(t)]$, $t = 1, \ldots, KM$. We take the DFT of the signals and form (7.144) for a fixed noise model $H^*(i\omega)$. Show that this gives exactly the same results as if we just take the DFT over one period and use the averaged output (7.153). Is it essential that the noise model is fixed?

7T.1 Suppose that a true description of a certain system is given by

$$y(t) + a_1^0 y(t-1) + \ldots + a_{n_a}^0 y(t-n_a) = b_1^0 u(t-1) + \ldots + b_{n_b}^0 u(t-n_b) + v_0(t)$$

for a stationary process $\{v_0(t)\}$ independent of the input. Let $\varphi(t)$ be defined, as usual, by (7.32), and let $\tilde{\varphi}(t)$ be given by

$$\tilde{\varphi}(t) = [-y_0(t-1) \ldots -y_0(t-n_a) \quad u(t-1) \ldots u(t-n_b)]^T$$

where

$$y_0(t) + a_1^0 y_0(t-1) + \ldots + a_{n_a}^0 y_0(t-n_a) = b_1^0 u(t-1) + \ldots + b_{n_b}^0 u(t-n_b)$$

Prove that for any vector of instrumental variables of the general kind (7.122) we have

$$E\zeta(t)\varphi^T(t) = E\zeta(t)\tilde{\varphi}^T(t)$$

7D.1 Consider the ARX structure (4.7) where one parameter, say b_1, is known to have a certain value b_1^*. Show that the associated predictor can be written as

$$\hat{y}(t|\theta) = \theta^T\varphi(t) + \mu(t)$$

with proper definitions of θ, φ, and μ (φ and μ to be known variables at time t). Derive the LS estimate and the IV estimate for this model.

7D.2 Let A be a given, positive symmetric definite matrix and let B and C be given matrices. Establish that

$$\theta^T A\theta - \theta^T B - B^T\theta + C = [\theta - A^{-1}B]^T A[\theta - A^{-1}B] + C - B^T A^{-1}B$$

$$\geq C - B^T A^{-1}B$$

and use this result to prove all the expressions for the LSE in Section 7.3 [(7.34), (7.41), (7.43), and (7.46)]. The matrix inequality $D \geq B$ is to be interpreted as "$D - B$ is a positive semidefinite matrix."

Hint: For (7.46), rewrite (7.45) as

$$V_N(\theta, Z^N) = \operatorname{tr}\frac{1}{N}\sum_{t=1}^{N}\left[y(t) - \theta^T\varphi(t)\right]\left[y(t) - \theta^T\varphi(t)\right]^T$$

7D.3 Let Σ be an invertible square $p \times p$ matrix with elements σ_{ij}. Prove the differentiation formula

$$\frac{\partial}{\partial\alpha_{ij}}\det\Sigma = \det[\Sigma]\cdot\mu_{ij}$$

where μ_{ij} is the i, j element of Σ^{-1}. [*Hint:* Use $\det(I + \varepsilon A) = 1 + \varepsilon \operatorname{tr} A +$ higher-order terms in ε]. Use the result to prove (7.94) and (7.95).

7D.4 Show that the two instrumental variable vectors, of dimension d, $\zeta_1(t)$ and $\zeta_2(t)$, where $\rho_1(t) = T\rho_2(t)$ with T invertible, give the same estimate $\hat{\theta}_N^{\mathrm{IV}}$ in (7.118).

7D.5 Show that if two variables x and u are associated as in (7.123) and (7.124) then we can write

$$\begin{bmatrix} -x(t-1) \\ \vdots \\ -x(t-n_n) \\ u(t-1) \\ \vdots \\ u(t-n_m) \end{bmatrix} = S(-M, N) \cdot \frac{1}{N(q)} \begin{bmatrix} u(t-1) \\ u(t-2) \\ \vdots \\ u(t-n_n-n_m) \end{bmatrix}$$

for an $(n_n + n_m) \times (n_n + n_m)$ matrix

$$S(-M, N) = \begin{bmatrix} -m_0 & -m_1 & \dots & -m_{n_m} & 0 & \dots & 0 \\ 0 & -m_0 & \dots & -m_{n_m-1} & -m_{n_m} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & -m_0 & -m_1 & \dots & -m_{n_m} \\ 1 & n_1 & \dots & n_{n_n} & 0 & \dots & 0 \\ 0 & 1 & \dots & n_{n_n-1} & n_{n_n} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & n_1 & \dots & n_{n_n} \end{bmatrix}$$

Such a matrix is called a *Sylvester matrix* (see, e.g., Kailath, 1980), and it will be nonsingular if and only if the polynomials in (7.124) have no common factor. Use this result to prove that the instruments (7.126) give the same IV estimate as the instruments (7.122). Reference: Söderström and Stoica (1983).

7D.6 Show that the prediction-error estimate obtained from (7.11) and (7.12) can also be seen as a correlation estimate (7.110) for a particular choice of L, ζ, and α.

7D.7 Give an explicit expression for the estimate $\hat{\theta}_N^{\text{EIV}}$ in (7.130) in the case ζ does not depend on θ, and $\alpha(\varepsilon) = \varepsilon$.

7D.8 Consider the symmetric matrix

$$H = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

Show that if $H \geq 0$, then

$$A - BC^{-1}B^T \geq 0.$$

Hint: Consider xHx^T for

$$x = [x_1 \quad -x_1 BC^{-1}]$$

with x_1 arbitrary.

APPENDIX 7A: PROOF OF THE CRAMÉR-RAO INEQUALITY

The assumption $E\hat{\theta}(y^N) = \theta_0$ can be written

$$\theta_0 = \int_{\mathbf{R}^N} \hat{\theta}(x^N) f_y(\theta_0, x^N)\, dx^N \tag{7A.1}$$

By definition we also have

$$1 = \int_{\mathbf{R}^N} f_y(\theta_0, x^N)\, dx^N \tag{7A.2}$$

Differentiating these two expressions with respect to θ_0 gives

$$\begin{aligned} I &= \int_{\mathbf{R}^N} \hat{\theta}(x^N) \left[\frac{d}{d\theta_0} f_y(\theta_0, x^N)\right]^T dx^N \\ &= \int_{\mathbf{R}^N} \hat{\theta}(x^N) \left[\frac{d}{d\theta_0} \log f_y(\theta_0, x^N)\right]^T f_y(\theta_0, x^N)\, dx^N \\ &= E\hat{\theta}(y^N) \left[\frac{d}{d\theta_0} \log f_y(\theta_0, y^N)\right]^T \end{aligned} \tag{7A.3}$$

(I is the $d \times d$ unit matrix) and

$$\begin{aligned} 0 &= \int_{\mathbf{R}^N} \left[\frac{d}{d\theta_0} f_y(\theta_0, x^N)\right]^T dx^N = \int_{\mathbf{R}^N} \left[\frac{d}{d\theta_0} \log f_y(\theta_0, x^N)\right]^T f_y(\theta_0, x^N) dx^N \\ &= E\left[\frac{d}{d\theta_0} \log f_y(\theta_0, y^N)\right]^T \end{aligned} \tag{7A.4}$$

Expectation in these two expressions is hence w.r.t. y^N.

Now multiply (7A.4) by θ_0 and subtract it from (7A.3). This gives

$$E\left[\hat{\theta}(y^N) - \theta_0\right] \left[\frac{d}{d\theta_0} \log f_y(\theta_0, y^N)\right]^T = I \tag{7A.5}$$

Now denote

$$\alpha = \hat{\theta}(y^N) - \theta_0, \qquad \beta = \frac{d}{d\theta_0} \log f_y(\theta_0, y^N) \tag{7A.6}$$

(both d-dimensional column vectors) so that

$$E\alpha\beta^T = I \tag{7A.7}$$

Hence

$$E\begin{bmatrix}\alpha \\ \beta\end{bmatrix}\begin{bmatrix}\alpha \\ \beta\end{bmatrix}^T = \begin{bmatrix} E\alpha\alpha^T & I \\ I & E\beta\beta^T \end{bmatrix} \geq 0$$

where the positive semidefiniteness follows by construction. Hence Problem 7D.8 proves that

$$E\alpha\alpha^T \geq [E\beta\beta^T]^{-1}$$

which is (7.79). It only remains to prove the equality in (7.80). Differentiating the transpose of (7A.4) gives

$$0 = \int_{\mathbf{R}^N} \left[\frac{d^2}{d\theta_0^2} \log f_y(\theta_0, x^N)\right] f_y(\theta_0, x^N)\, dx^N$$

$$+ \int_{\mathbf{R}^N} \left[\frac{d}{d\theta_0} \log f_y(\theta_0, x^N)\right]\left[\frac{d}{d\theta_0} \log f_y(\theta_0, x^N)\right]^T f_y(\theta_0, x^N)\, dx^N$$

which gives (7.80).