# Appendix C

# Model Order Selection Tools

## C.1 INTRODUCTION

The parametric methods of spectral analysis (discussed in Chapters 3, 4, and 6) require not only the estimation of a vector of real-valued parameters but also the selection of one or several *integer-valued* parameters that are equally important for the specification of the data model. Specifically, these integer-valued parameters of the model are the ARMA model orders (in Chapter 3), the number of sinusoidal components (in Chapter 4), and the number of source signals impinging on the array (in Chapter 6). In each of these cases, the integer-valued parameters determine the dimension of the real-valued parameter vector of the data model. In what follows, we will use the following symbols:

$$y = \text{the vector of available data (of size } N)$$

$$\theta = \text{the (real-valued) parameter vector}$$

$$n = \text{the dimension of } \theta$$

For short, we will refer to $n$ as the *model order*, even though sometimes $n$ is not really an order. (See, for example, the preceding examples.) We assume that both $y$ and $\theta$ are real valued:

$$y \in \mathbf{R}^N, \qquad \theta \in \mathbf{R}^n$$

Whenever we need to emphasize that the number of elements in $\theta$ is $n$, we will use the notation $\theta^n$. A method that estimates $n$ from the data vector $y$ will be called an *order-selection rule*. Note

that the need for estimating a model order is typical of the parametric approaches to spectral analysis. The nonparametric methods of spectral analysis do not have such a requirement.

The discussion in the text on the parametric spectral methods has focused on estimating the model-parameter vector $\theta$ for a specific order $n$. In this general appendix (based on [STOICA and SELÉN 2004b]) we explain how to estimate $n$ as well. The literature on order selection is as considerable as that on (real-valued) parameter estimation (see, e.g., [CHOI 1992; SÖDERSTRÖM AND STOICA 1989; MCQUARRIE AND TSAI 1998; LINHART AND ZUCCHINI 1986; BURNHAM AND ANDERSON 2002; SAKAMOTO, ISHIGURO, AND KITAGAWA 1986; STOICA, EYKHOFF, JANNSEN, AND SÖDERSTRÖM 1986] and the many references therein). However, many order selection rules are tied to specific parameter estimation methods; hence, their applicability is rather limited. Here we will concentrate on order-selection rules that are associated with the maximum likelihood method (MLM) of parameter estimation. As explained briefly in Appendix B (and in what follows here), the MLM is likely the most commonly used parameter estimation method. Consequently, the order estimation rules that can be used with the MLM are of quite a general interest. In the next section, we review briefly the ML method of parameter estimation and some of its main properties.

## C.2 MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

Let

$p(y, \theta)$ = the probability density function (pdf) of the data vector $y$, which depends on the parameter vector $\theta$; also called *the likelihood function*.

The ML estimate of $\theta$, which we denote by $\hat{\theta}$, is given by the maximizer of $p(y, \theta)$ (see, for example, [ANDERSON 1971; BROCKWELL AND DAVIS 1991; HANNAN AND DEISTLER 1988; PAPOULIS 1977; PORAT 1994; PRIESTLEY 1981; SCHARF 1991; THERRIEN 1992; SÖDERSTRÖM AND STOICA 1989] and Appendix B). Alternatively, because $\ln(\cdot)$ is a monotonically increasing function,

$$\hat{\theta} = \arg\max_{\theta} \ln p(y, \theta) \tag{C.2.1}$$

Under the Gaussian data assumption, the MLM typically reduces to the nonlinear least-squares (NLS) method of parameter estimation (particular forms of which are discussed briefly in Chapter 3 and in more detail in Chapters 4 and 6). To illustrate this fact, let us assume that the observation vector $y$ can be written as

$$y = \mu(\gamma) + e \tag{C.2.2}$$

where $e$ is a (real-valued) Gaussian white-noise vector with mean zero and covariance matrix given by $E\left\{ee^T\right\} = \sigma^2 I$, $\gamma$ is an unknown parameter vector, and $\mu(\gamma)$ is a deterministic function

of $\gamma$. It follows readily from (C.2.2) that

$$p(y, \theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{\|y-\mu(\gamma)\|^2}{2\sigma^2}} \tag{C.2.3}$$

where

$$\theta = \begin{bmatrix} \gamma \\ \sigma^2 \end{bmatrix} \tag{C.2.4}$$

**Remark:** Note that, in this appendix, we use the symbol $\theta$ for the whole parameter vector, unlike in some previous discussions, where we used $\theta$ to denote the signal parameter vector (which is denoted by $\gamma$ here). ■

We deduce from (C.2.3) that

$$-2 \ln p(y, \theta) = N \ln(2\pi) + N \ln \sigma^2 + \frac{\|y - \mu(\gamma)\|^2}{\sigma^2} \tag{C.2.5}$$

A simple calculation based on (C.2.5) shows that the ML estimates of $\gamma$ and $\sigma^2$ are given by

$$\hat{\gamma} = \arg \min_{\gamma} \|y - \mu(\gamma)\|^2 \tag{C.2.6}$$

$$\hat{\sigma}^2 = \frac{1}{N} \|y - \mu(\hat{\gamma})\|^2 \tag{C.2.7}$$

The corresponding value of the likelihood function is given by

$$\boxed{-2 \ln p(y, \hat{\theta}) = \text{constant} + N \ln \hat{\sigma}^2} \tag{C.2.8}$$

As can be seen from (C.2.6), in the present case the MLM indeed reduces to the NLS. In particular, note that the NLS method for sinusoidal parameter estimation discussed in Chapter 4 is precisely of the form of (C.2.6). If we let $N_s$ denote the number of observed complex-valued samples of the noisy sinusoidal signal and $n_c$ denote the number of sinusoidal components present in the signal, then

$$N = 2N_s \tag{C.2.9}$$

$$n = 3n_c + 1 \tag{C.2.10}$$

We will use the sinusoidal signal model of Chapter 4 as a vehicle for illustrating how the various general order-selection rules presented in what follows should be used in a specific situation. These rules can also be used with the parametric spectral analysis methods of Chapters 3 and 6. The task of deriving *explicit forms* of these order selection rules for the aforementioned methods is left as an interesting exercise to the reader (see, for example, [McQuarrie and Tsai 1998; Brockwell and Davis 1991; Porat 1994]).

Next, we note that, under regularity conditions, the pdf of the ML estimate $\hat{\theta}$ converges, as $N \to \infty$, to a Gaussian pdf with mean $\theta$ and covariance matrix equal to the Cramér–Rao bound (CRB) matrix (see Section B.2 for a discussion about the CRB). Consequently, asymptotically in $N$, the pdf of $\hat{\theta}$ is given by

$$p(\hat{\theta}) = \frac{1}{(2\pi)^{n/2}|J^{-1}|^{1/2}} e^{-\frac{1}{2}(\hat{\theta}-\theta)^T J (\hat{\theta}-\theta)} \tag{C.2.11}$$

where (see (B.2.10))

$$J = -E\left\{ \frac{\partial^2 \ln p(y,\theta)}{\partial\theta\ \partial\theta^T} \right\} \tag{C.2.12}$$

**Remark:** To simplify the notation, we use the symbol $\theta$ for both the true parameter vector and the parameter vector viewed as an unknown variable (as we also did in Appendix B). The exact meaning of $\theta$ should be clear from the context. ■

The "regularity conditions" referred to previously require that $n$ not be a function of $N$ and, hence, that the ratio between the number of unknown parameters and the number of observations tends to zero as $N \to \infty$. This is true for the parametric spectral analysis problems discussed in Chapters 3 and 4. However, the previous condition does not hold for the parametric spectral analysis problem addressed in Chapter 6. Indeed, in the latter case, the number of parameters to be estimated from the data is proportional to $N$, because the signal sequence is completely unknown. To overcome this difficulty, we can assume that the signal vector is temporally white and Gaussian distributed, which leads to a ML problem that satisfies the previously stated regularity condition. (We refer the interested reader to [Ottersten, Viberg, Stoica, and Nehorai 1993; Stoica and Nehorai 1990; Van Trees 2002] for details on this ML approach to the spatial spectral analysis problem of Chapter 6.)

To close this section, we note that, under mild conditions,

$$\left[ -\frac{1}{N} \frac{\partial^2 \ln p(y,\theta)}{\partial\theta\ \partial\theta^T} - \frac{1}{N} J \right] \to 0 \quad \text{as } N \to \infty \tag{C.2.13}$$

To motivate (C.2.13) for the fairly general data model in (C.2.2), we can argue as follows: Let us rewrite the negative log-likelihood function associated with (C.2.2) (see (C.2.5)) as

$$-\ln p(y,\theta) = \text{constant} + \frac{N}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]^2 \tag{C.2.14}$$

where the subindex $t$ denotes the $t$-th component. From (C.2.14), we obtain, by a simple calculation,

$$-\frac{\partial \ln p(y,\theta)}{\partial\theta} = \begin{bmatrix} -\dfrac{1}{\sigma^2}\displaystyle\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]\mu_t'(\gamma) \\ \dfrac{N}{2\sigma^2} - \dfrac{1}{2\sigma^4}\displaystyle\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]^2 \end{bmatrix} \tag{C.2.15}$$

where

$$\mu_t'(\gamma) = \frac{\partial \mu_t(\gamma)}{\partial \gamma} \tag{C.2.16}$$

Differentiating (C.2.15) once again gives

$$\begin{aligned} &-\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\ \partial\theta^T} \\ &= \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{t=1}^{N} e_t \mu_t''(\gamma) + \frac{1}{\sigma^2}\sum_{t=1}^{N} \mu_t'(\gamma)\mu_t'^T(\gamma) & \frac{1}{\sigma^4}\sum_{t=1}^{N} e_t \mu_t'(\gamma) \\ \frac{1}{\sigma^4}\sum_{t=1}^{N} e_t \mu_t'(\gamma) & -\frac{N}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{t=1}^{N} e_t^2 \end{bmatrix} \end{aligned} \tag{C.2.17}$$

where $e_t = y_t - \mu_t(\gamma)$ and

$$\mu_t''(\gamma) = \frac{\partial^2 \mu_t(\gamma)}{\partial\gamma\ \partial\gamma^T} \tag{C.2.18}$$

Taking the expectation of (C.2.17) and dividing by $N$, we get

$$\frac{1}{N}J = \begin{bmatrix} \frac{1}{\sigma^2}\left(\frac{1}{N}\sum_{t=1}^{N} \mu_t'(\gamma)\mu_t'^T(\gamma)\right) & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \tag{C.2.19}$$

We assume that $\mu(\gamma)$ is such that the previous matrix has a finite limit as $N \to \infty$. Under this assumption and the previously-made assumption on $e$, we can also show from (C.2.17) that

$$-\frac{1}{N}\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\ \partial\theta^T}$$

converges (as $N \to \infty$) to the right side of (C.2.19), which concludes the motivation of (C.2.13). Letting

$$\hat{J} = -\left.\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\ \partial\theta^T}\right|_{\theta=\hat{\theta}} \tag{C.2.20}$$

we deduce from (C.2.13), (C.2.19), and the consistency of $\hat{\theta}$ that, for sufficiently large values of $N$,

$$\frac{1}{N}\hat{J} \simeq \frac{1}{N}J = \mathcal{O}(1) \tag{C.2.21}$$

Hereafter, $\simeq$ denotes an asymptotic (approximate) equality, in which the higher order terms have been neglected, and $\mathcal{O}(1)$ denotes a term that tends to a constant as $N \to \infty$.

Interestingly enough, the assumption that the right side of (C.2.19) has a finite limit, as $N \to \infty$, holds for many problems, but *not* for the sinusoidal parameter estimation problem of Chapter 4. In the latter case, (C.2.21) needs to be modified to (see, e.g., Appendix B)

$$K_N \hat{J} K_N \simeq K_N J K_N = \mathcal{O}(1) \tag{C.2.22}$$

where

$$K_N = \begin{bmatrix} \frac{1}{N_s^{3/2}} I_{n_c} & 0 \\ 0 & \frac{1}{N_s^{1/2}} I_{2n_c+1} \end{bmatrix} \tag{C.2.23}$$

and where $I_k$ denotes the $k \times k$ identity matrix; to write (C.2.23), we assumed that the upper left $n_c \times n_c$ block of $J$ corresponds to the sinusoidal frequencies, but this fact is not really important for the analysis in this appendix, as we will see below.

## C.3 USEFUL MATHEMATICAL PRELIMINARIES AND OUTLOOK

In this section, we discuss a number of mathematical tools that will be used in the next sections to derive several important order-selection rules. We will keep the discussion at an informal level to make the material as accessible as possible. In Section C.3.1, we will formulate the model order selection as a hypothesis-testing problem, with the main goal of showing that the maximum *a posteriori* (MAP) approach leads to the optimal order-selection rule (in a sense specified there). In Section C.3.2, we discuss the Kullback–Leibler information criterion, which lies at the basis of another approach that can be used to derive model order selection rules.

### C.3.1 Maximum *A Posteriori* (MAP) Selection Rule

Let $H_n$ denote the hypothesis that the model order is $n$, and let $\bar{n}$ denote a known upper bound on $n$:

$$n \in [1, \bar{n}] \tag{C.3.1}$$

We assume that the hypotheses $\{H_n\}_{n=1}^{\bar{n}}$ are *mutually exclusive* (i.e., only one of them can hold true at a time). As an example, for a real-valued AR signal with coefficients $\{a_k\}$, we can define $H_n$ as follows:

$$H_n : \ a_n \neq 0 \text{ and } a_{n+1} = \cdots = a_{\bar{n}} = 0 \tag{C.3.2}$$

For a sinusoidal signal we can proceed similarly, after observing that, for such a signal, the number of components $n_c$ is related to $n$ as in (C.2.10), *viz.*,

$$n = 3n_c + 1 \tag{C.3.3}$$

Hence, for a sinusoidal signal with amplitudes $\{\alpha_k\}$, we can consider the following hypotheses:

$$H_{n_c}: \ \alpha_k \neq 0 \text{ for } k = 1, \ldots, n_c, \text{ and } \alpha_k = 0 \text{ for } k = n_c + 1, \ldots, \bar{n}_c \tag{C.3.4}$$

for $n_c \in [1, \bar{n}_c]$ (with the corresponding "model order" $n$ being given by (C.3.3)).

**Remark:** The hypotheses $\{H_n\}$ can be *either nested or non-nested*. We say that $H_1$ and $H_2$ are nested whenever the model corresponding to $H_1$ can be obtained as a special case of that associated with $H_2$. To give an example, the following hypotheses:

$H_1$ : the signal is a first-order AR process

$H_2$ : the signal is a second-order AR process

are nested, whereas the $H_1$ and

$H_3$ : the signal consists of one sinusoid in noise

are nonnested. ■

Let

$$p_n(y|H_n) = \text{the pdf of } y \text{ under } H_n \tag{C.3.5}$$

Whenever we want to emphasize the possible dependence of the pdf in (C.3.5) on the parameter vector of the model corresponding to $H_n$, we write

$$p_n(y, \theta^n) \triangleq p_n(y|H_n) \tag{C.3.6}$$

Assuming that (C.3.5) is available, along with the *a priori* probability of $H_n$, $p_n(H_n)$, we can write the conditional probability of $H_n$, given $y$, as

$$p_n(H_n|y) = \frac{p_n(y|H_n)p_n(H_n)}{p(y)} \tag{C.3.7}$$

The maximum *a posteriori* probability (MAP) rule selects the order $n$ (or the hypothesis $H_n$) that maximizes (C.3.7). The denominator in (C.3.7) does not depend on $n$, so the *MAP rule* is given by

$$\boxed{\max_{n \in [1, \bar{n}]} p_n(y|H_n)p_n(H_n)} \tag{C.3.8}$$

Most typically, the hypotheses $\{H_n\}$ are *a priori equiprobable*—that is,

$$p_n(H_n) = \frac{1}{\bar{n}}, \quad n = 1, \ldots, \bar{n} \tag{C.3.9}$$

In such a case the MAP rule reduces to

$$\max_{n\in[1,\bar{n}]} p_n(y|H_n) \tag{C.3.10}$$

Next, we define the *average (or total) probability of correct detection* as

$$P_{cd} = \Pr\{[(\text{decide } H_1) \cap (H_1 = \text{true})] \cup \cdots \cup [(\text{decide } H_{\bar{n}}) \cap (H_{\bar{n}} = \text{true})]\} \tag{C.3.11}$$

The attribute "average" that has been attached to $P_{cd}$ is motivated by the fact that (C.3.11) gives the probability of correct detection "averaged" over all possible hypotheses (as opposed, for example, to only considering the probability of correctly detecting that the model order is 2 (let us say), which is $\Pr\{\text{decide } H_2|H_2\}$).

**Remark:** Regarding the terminology, note that the determination of a real-valued parameter from the available data is called "estimation," whereas it is usually called "detection" for an integer-valued parameter, such as a model order. ■

In the following, we prove that *the MAP rule is optimal in the sense of maximizing* $P_{cd}$. To do so, consider a generic rule for selecting $n$, or, equivalently, for testing the hypotheses $\{H_n\}$ against each other. Such a rule will implicitly or explicitly partition the observation space, $\mathbf{R}^N$, into $\bar{n}$ sets $\{S_n\}_{n=1}^{\bar{n}}$, which are such that

$$\text{We decide } H_n \text{ if and only if } y \in S_n \tag{C.3.12}$$

Making use of (C.3.12) along with the fact that the hypotheses $\{H_n\}$ are mutually exclusive, we can write $P_{cd}$ in (C.3.11) as

$$\begin{aligned} P_{cd} &= \sum_{n=1}^{\bar{n}} \Pr\{(\text{decide } H_n) \cap (H_n = \text{true})\} \\ &= \sum_{n=1}^{\bar{n}} \Pr\{(\text{decide } H_n)|H_n\} \Pr\{H_n\} \\ &= \sum_{n=1}^{\bar{n}} \int_{S_n} p_n(y|H_n) p_n(H_n)\, dy \\ &= \int_{\mathbf{R}^N} \left[ \sum_{n=1}^{\bar{n}} I_n(y) p_n(y|H_n) p_n(H_n) \right] dy \end{aligned} \tag{C.3.13}$$

where $I_n(y)$ is the so-called indicator function, given by

$$I_n(y) = \begin{cases} 1, & \text{if } y \in S_n \\ 0, & \text{otherwise} \end{cases} \tag{C.3.14}$$

Next, observe that, for any given data vector, $y$, one and only one indicator function can be equal to 1 (because the sets $S_n$ do not overlap, and their union is $\mathbf{R}^N$). This observation, along with the expression (C.3.13) for $P_{cd}$, implies that the MAP rule in (C.3.8) maximizes $P_{cd}$, as stated. Note that the sets $\{S_n\}$ corresponding to the MAP rule are implicitly defined via (C.3.8); however, $\{S_n\}$ are of no real interest in the proof, as both they and the indicator functions are introduced only to simplify the above proof. For more details on the topic of this subsection, we refer the reader to [Scharf 1991; Van Trees 1968].

### C.3.2 Kullback–Leibler Information

Let $p_0(y)$ denote the *true pdf* of the observed data vector $y$, and let $\hat{p}(y)$ denote the pdf of a generic model of the data. The "discrepancy" between $p_0(y)$ and $\hat{p}(y)$ can be measured by using the Kullback–Leibler (KL) information or discrepancy function (see [Kullback and Leibler 1951]):

$$D(p_0, \hat{p}) = \int p_0(y) \ln \left[ \frac{p_0(y)}{\hat{p}(y)} \right] dy \tag{C.3.15}$$

To simplify the notation, we omit the region of integration when it is the entire space. Letting $E_0\{\cdot\}$ denote the expectation with respect to the true pdf, $p_0(y)$, we can rewrite (C.3.15) as

$$D(p_0, \hat{p}) = E_0 \left\{ \ln \left[ \frac{p_0(y)}{\hat{p}(y)} \right] \right\} = E_0\{\ln p_0(y)\} - E_0\{\ln \hat{p}(y)\} \tag{C.3.16}$$

Next, we prove that (C.3.15) possesses some properties of a suitable discrepancy function—namely,

$$\boxed{\begin{aligned} &D(p_0, \hat{p}) \geq 0 \\ &D(p_0, \hat{p}) = 0 \text{ if and only if } p_0(y) = \hat{p}(y) \end{aligned}} \tag{C.3.17}$$

To verify (C.3.17), we use the fact shown in Complement 6.5.8, that

$$-\ln \lambda \geq 1 - \lambda \quad \text{for any } \lambda > 0 \tag{C.3.18}$$

and

$$-\ln \lambda = 1 - \lambda \quad \text{if and only if } \lambda = 1 \tag{C.3.19}$$

Hence, letting $\lambda(y) = \hat{p}(y)/p_0(y)$, we have that

$$D(p_0, \hat{p}) = \int p_0(y) \left[ -\ln \lambda(y) \right] dy$$
$$\geq \int p_0(y) \left[ 1 - \lambda(y) \right] dy = \int p_0(y) \left[ 1 - \frac{\hat{p}(y)}{p_0(y)} \right] dy = 0$$

where the equality holds if and only if $\lambda(y) \equiv 1$, i.e. $\hat{p}(y) \equiv p_0(y)$.

**Remark:** The inequality in (C.3.17) also follows from Jensen's inequality (see equation (4.9.36) in Complement 4.9.5) and from the concavity of the function $\ln(\cdot)$:

$$D(p_0, \hat{p}) = -E_0 \left\{ \ln \left[ \frac{\hat{p}(y)}{p_0(y)} \right] \right\}$$
$$\geq -\ln \left[ E_0 \left\{ \frac{\hat{p}(y)}{p_0(y)} \right\} \right]$$
$$= -\ln \left[ \int \frac{\hat{p}(y)}{p_0(y)} p_0(y)\, dy \right] = -\ln(1) = 0 \qquad \blacksquare$$

The KL discrepancy function can be viewed as quantifying the "loss of information" induced by the use of $\hat{p}(y)$ in lieu of $p_0(y)$. For this reason, $D(p_0, \hat{p})$ is sometimes called an information function, and the order-selection rules derived from it are called *information criteria* (see Sections C.4–C.6).

### C.3.3 Outlook: Theoretical and Practical Perspectives

Neither the MAP rule nor the KL information can be used directly for order selection, because neither the pdfs of the data vector under the various hypotheses nor the true data pdf are available in any of the parametric spectral analysis problems discussed in the text. A possible way of using the MAP approach for order estimation consists of assuming an *a priori* pdf for the unknown parameter vector, $\theta^n$, and integrating $\theta^n$ out of $p_n(y, \theta^n)$ to obtain $p_n(y|H_n)$. This Bayesian-type approach will be discussed in Section C.7. Regarding the KL approach, a natural way of using it for order selection consists in using an estimate, $\hat{D}(p_0, \hat{p})$, in lieu of the unavailable $D(p_0, \hat{p})$ (for a suitably chosen model pdf, $\hat{p}(y)$), and in determining the model order by minimizing $\hat{D}(p_0, \hat{p})$. This KL-based approach will be discussed in Sections C.4–C.6.

The derivations of all model order selection rules in the sections that follow rely on the assumption that one of the hypotheses $\{H_n\}$ is true. This assumption is unlikely to hold in applications with real-life data, so the reader will justifiably wonder whether an order-selection rule derived under such an assumption has any practical value. To address this concern, we remark that good parameter estimation methods (such as the MLM), derived under rather strict modeling assumptions, perform quite well in applications where the assumptions made are rarely satisfied exactly. Similarly, order-selection rules based on sound theoretical principles (such as the ML, KL, and MAP principles used in this text) are likely to perform well in applications despite the fact

that some of the assumptions made when deriving them do not hold exactly. The precise behavior of order-selection rules (such as those presented in the sections to follow) in various mismodeling scenarios is not well understood, but extensive simulation results (see, e.g., [McQuarrie and Tsai 1998; Linhart and Zucchini 1986; Burnham and Anderson 2002]) lend support to this claim.

## C.4 DIRECT KULLBACK–LEIBLER (KL) APPROACH: NO-NAME RULE

The model-dependent part of the Kullback–Leibler (KL) information, (C.3.16), is given by

$$-E_0\{\ln \hat{p}(y)\} \tag{C.4.1}$$

where $\hat{p}(y)$ is the pdf or likelihood of the model (to simplify the notation, we omit the index $n$ of $\hat{p}(y)$; we will reinstate the index $n$ later on, when needed). Minimization of (C.4.1) with respect to the model order is equivalent to *maximization* of the function

$$I(p_0, \hat{p}) \triangleq E_0\{\ln \hat{p}(y)\} \tag{C.4.2}$$

which is sometimes called the relative KL information. The ideal choice for $\hat{p}(y)$ in (C.4.2) would be the model likelihood, $p_n(y|H_n) = p_n(y, \theta^n)$. However, the model likelihood function is not available, and hence this choice is not possible. Instead, we might think of using

$$\hat{p}(y) = p(y, \hat{\theta}) \tag{C.4.3}$$

in (C.4.2), which would give

$$I\left(p_0, p(y, \hat{\theta})\right) = E_0\left\{\ln p(y, \hat{\theta})\right\} \tag{C.4.4}$$

Because the true pdf of the data vector is unknown, we cannot evaluate the expectation in (C.4.4). Apparently, what we could easily do is use the following unbiased estimate of $I\left(p_0, p(y, \hat{\theta})\right)$, instead of (C.4.4) itself:

$$\hat{I} = \ln p(y, \hat{\theta}) \tag{C.4.5}$$

However, the order-selection rule that maximizes (C.4.5) does *not* have satisfactory properties. This is especially true for *nested models*, in the case of which the order-selection rule based on the maximization of (C.4.5) *fails completely*: indeed, for nested models, this rule will always choose the maximum possible order, $\bar{n}$, because $\ln p_n(y, \hat{\theta}^n)$ increases monotonically with increasing $n$.

A better idea consists of approximating the unavailable log-pdf of the model, $\ln p_n(y, \theta^n)$, by a second-order Taylor series expansion around $\hat{\theta}^n$, and then using the approximation so obtained to define $\ln \hat{p}(y)$ in (C.4.2):

$$\begin{aligned} \ln p_n(y, \theta^n) \simeq \ln p_n(y, \hat{\theta}^n) &+ (\theta^n - \hat{\theta}^n)^T \left[ \frac{\partial \ln p_n(y, \theta^n)}{\partial \theta^n} \bigg|_{\theta^n = \hat{\theta}^n} \right] \\ &+ \frac{1}{2}(\theta^n - \hat{\theta}^n)^T \left[ \frac{\partial^2 \ln p_n(y, \theta^n)}{(\partial \theta^n)\,(\partial \theta^n)^T} \bigg|_{\theta^n = \hat{\theta}^n} \right] (\theta^n - \hat{\theta}^n) \triangleq \ln \hat{p}_n(y) \end{aligned} \tag{C.4.6}$$

Because $\hat{\theta}^n$ is the maximizer of $\ln p_n(y, \theta^n)$, the second term in (C.4.6) is equal to zero. Hence, we can write (see also (C.2.21))

$$\ln \hat{p}_n(y) \simeq \ln p_n(y, \hat{\theta}^n) - \frac{1}{2}(\theta^n - \hat{\theta}^n)^T J (\theta^n - \hat{\theta}^n) \tag{C.4.7}$$

According to (C.2.11),

$$E_0\left\{(\theta^n - \hat{\theta}^n)^T J (\theta^n - \hat{\theta}^n)\right\} = \text{tr}\left[J E_0\left\{(\theta^n - \hat{\theta}^n)(\theta^n - \hat{\theta}^n)^T\right\}\right] = \text{tr}[I_n] = n \tag{C.4.8}$$

which means that, for the choice of $\hat{p}_n(y)$ in (C.4.7), we have

$$I = E_0\left\{\ln p_n(y, \hat{\theta}^n) - \frac{n}{2}\right\} \tag{C.4.9}$$

An unbiased estimate of the above relative KL information is given by

$$\ln p_n(y, \hat{\theta}^n) - \frac{n}{2} \tag{C.4.10}$$

The corresponding order-selection rule maximizes (C.4.10), or, equivalently, *minimizes*

$$\text{NN}(n) = -2 \ln p_n(y, \hat{\theta}^n) + n \tag{C.4.11}$$

with respect to model order $n$. This no-name (NN) rule can be shown to perform better than that based on (C.4.5), but worse than the rules presented in the next sections. Essentially, the problem with (C.4.11) is that it tends to overfit (i.e., to select model orders larger than the "true" order). To understand intuitively how this happens, note that the first term in (C.4.11) decreases with increasing $n$ (for nested models), whereas the second term increases. Hence, the second term in (C.4.11) *penalizes overfitting*; however, it turns out that it does not penalize quite enough. The rules presented in the following sections have a form similar to (C.4.11), but with a larger penalty term, and they do have better properties than (C.4.11). Despite this fact, we have chosen to present (C.4.11) briefly in this section for two reasons: (i) the discussion here has revealed the failure of using $\max_n \ln p_n(y, \hat{\theta}^n)$ as an order-selection rule *and* has shown that it is in effect quite easy to obtain rules with better properties; and (ii) this section has laid groundwork for the derivation of better order-selection rules based on the KL approach in the next two sections.

To close this section, we motivate the multiplication by $-2$ in going from (C.4.10) to (C.4.11). The reason for preferring (C.4.11) to (C.4.10) is that for the fairly common NLS model in (C.2.2) and the associated Gaussian likelihood in (C.2.3), $-2 \ln p_n(y, \hat{\theta}^n)$ takes on the following convenient form:

$$-2 \ln p_n(y, \hat{\theta}^n) = N \ln \hat{\sigma}_n^2 + \text{constant} \tag{C.4.12}$$

(See (C.2.5)–(C.2.7).) Hence, in such a case, we can replace $-2 \ln p_n(y, \hat{\theta}^n)$ in (C.4.11) by the scaled logarithm of the residual variance, $N \ln \hat{\sigma}_n^2$. This remark also applies to the order-selection rules presented in the following sections, which are written in a form similar to (C.4.11).

## C.5 CROSS-VALIDATORY KL APPROACH: THE AIC RULE

As explained in the previous section, a possible approach to model order selection consists of minimizing the KL discrepancy between the "true" pdf of the data and the pdf (or likelihood) of the model, or, equivalently, of maximizing the relative KL information (see (C.4.2)):

$$I(p_0, \hat{p}) = E_0\{\ln \hat{p}(y)\} \tag{C.5.1}$$

When using this approach, the first (and, likely the main) hurdle that we have to overcome is *the choice of the model likelihood*, $\hat{p}(y)$. As discussed in the previous section, we would ideally like to use the true pdf of the model as $\hat{p}(y)$ in (C.5.1), i.e. $\hat{p}(y) = p_n(y, \theta^n)$, but this is not possible; $p_n(y, \theta^n)$ is unknown. Hence, we have to choose $\hat{p}(y)$ in a different way. This choice is important; it eventually determines the model order selection rule that we will obtain. The other issue we should consider when using the approach based on (C.5.1) is that *the expectation in* (C.5.1) *cannot be evaluated*, because the true pdf of the data is unknown. Consequently, we will have to use an estimate, $\hat{I}$, in lieu of the unavailable $I(p_0, \hat{p})$ in (C.5.1).

Let $x$ denote a *fictitious* data vector having the same size, $N$, and the same pdf as $y$, but such that $x$ is *independent* of $y$. Also, let $\hat{\theta}_x$ denote the ML estimate of the model parameter vector that would be obtained from $x$ if $x$ were available. (We omit the superindex $n$ of $\hat{\theta}_x$ as often as possible, to simplify notation.) In this section, we will consider the following choice of the model's pdf:

$$\ln \hat{p}(y) = E_x\left\{\ln p(y, \hat{\theta}_x)\right\} \tag{C.5.2}$$

which, when inserted in (C.5.1), yields

$$I = E_y\left\{E_x\left\{\ln p(y, \hat{\theta}_x)\right\}\right\} \tag{C.5.3}$$

Hereafter, $E_x\{\cdot\}$ and $E_y\{\cdot\}$ denote the expectation with respect to the pdf of $x$ and $y$, respectively. The above choice of $\hat{p}(y)$, which was introduced in [Akaike 1974; Akaike 1978], has an interesting *cross-validation interpretation*: we use the sample $x$ for estimation and the independent sample $y$ for validation of the estimated model's pdf. Note that the dependence of (C.5.3) on the fictitious sample $x$ is eliminated (as it should be, because $x$ is unavailable) via the expectation operation $E_x\{\cdot\}$; see below for details.

An asymptotic second-order Taylor series expansion of $\ln p(y, \hat{\theta}_x)$ around $\hat{\theta}_y$, similar to (C.4.6)–(C.4.7), yields

$$\begin{aligned}
\ln p(y, \hat{\theta}_x) &\simeq \ln p(y, \hat{\theta}_y) + (\hat{\theta}_x - \hat{\theta}_y)^T \left[\left.\frac{\partial \ln p(y, \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_y}\right] \\
&\quad + \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)^T \left[\left.\frac{\partial^2 \ln p(y, \theta)}{\partial \theta\ \partial \theta^T}\right|_{\theta=\hat{\theta}_y}\right](\hat{\theta}_x - \hat{\theta}_y) \\
&\simeq \ln p(y, \hat{\theta}_y) - \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)^T J_y (\hat{\theta}_x - \hat{\theta}_y)
\end{aligned} \tag{C.5.4}$$

where $J_y$ is the $J$ matrix, as defined in (C.2.20), associated with the data vector $y$. Using the fact that $x$ and $y$ have the same pdf (which implies that $J_y = J_x$), along with the fact that they are independent of each other, we can show that

$$\begin{aligned} & E_y\left\{E_x\left\{(\hat{\theta}_x - \hat{\theta}_y)^T J_y (\hat{\theta}_x - \hat{\theta}_y)\right\}\right\} \\ & \quad = E_y\left\{E_x\left\{\mathrm{tr}\left(J_y\left[(\hat{\theta}_x - \theta) - (\hat{\theta}_y - \theta)\right]\left[(\hat{\theta}_x - \theta) - (\hat{\theta}_y - \theta)\right]^T\right)\right\}\right\} \\ & \quad = \mathrm{tr}\left[J_y\left(J_x^{-1} + J_y^{-1}\right)\right] = 2n \end{aligned} \tag{C.5.5}$$

Inserting (C.5.5) in (C.5.4) yields the following asymptotic approximation of the relative KL information in (C.5.3):

$$I \simeq E_y\left\{\ln p_n(y, \hat{\theta}^n) - n\right\} \tag{C.5.6}$$

(where we have omitted the subindex $y$ of $\hat{\theta}$ but reinstated the superindex $n$). Evidently, (C.5.6) can be estimated in an unbiased manner by

$$\ln p_n(y, \hat{\theta}^n) - n \tag{C.5.7}$$

Maximizing (C.5.7) with respect to $n$ is equivalent to *minimizing* the function of $n$

$$\boxed{\mathrm{AIC} = -2 \ln p_n(y, \hat{\theta}^n) + 2n} \tag{C.5.8}$$

where the acronym AIC stands for *Akaike Information Criterion* (the reasons for multiplying (C.5.7) by $-2$ to get (C.5.8), and for the use of the word "information" in the name given to (C.5.8) have been explained before—see the previous two sections).

As an example, for *the sinusoidal signal model* with $n_c$ components (see Section C.2), AIC takes on the form (see (C.2.6)–(C.2.10))

$$\mathrm{AIC} = 2N_s \ln \hat{\sigma}_{n_c}^2 + 2(3n_c + 1) \tag{C.5.9}$$

where $N_s$ denotes the number of available complex-valued samples, $\{y_c(t)\}_{t=1}^{N_s}$, and

$$\hat{\sigma}_{n_c}^2 = \frac{1}{N_s} \sum_{t=1}^{N_s} \left| y_c(t) - \sum_{k=1}^{n_c} \hat{\alpha}_k e^{i(\hat{\omega}_k t + \hat{\varphi}_k)} \right|^2 \tag{C.5.10}$$

**Remark:** AIC can also be obtained by using the following relative KL information function, in lieu of (C.5.3):

$$I = E_y\left\{E_x\left\{\ln p(x, \hat{\theta}_y)\right\}\right\} \tag{C.5.11}$$

Note that, in (C.5.11), $x$ is used for validation and $y$ for estimation. However, the derivation of AIC from (C.5.11) is more complicated; such a derivation, which is left as an exercise to the reader, makes use of two Taylor series expansions and of the fact that $E_x\{\ln p(x, \theta)\} = E_y\{\ln p(y, \theta)\}$.■

The performance of AIC has been found to be satisfactory in many case studies and applications to real-life data reported in the literature (see, for example, [McQuarrie and Tsai 1998; Linhart and Zucchini 1986; Burnham and Anderson 2002; Sakamoto, Ishiguro, and Kitagawa 1986]). *The performance of a model order selection rule*, such as AIC, can be measured in different ways, as explained in the next two paragraphs.

As a first possibility, we can consider a scenario in which the data-generating mechanism belongs to the class of models under test; thus, there is a true order. In such a case, analytical or numerical studies can be used to determine *the probability with which the rule selects the true order*. For AIC, it can be shown that, under quite general conditions,

$$\text{the probability of underfitting} \to 0 \tag{C.5.12}$$

$$\text{the probability of overfitting} \to \text{constant} > 0 \tag{C.5.13}$$

as $N \to \infty$ (see, for example, [McQuarrie and Tsai 1998; Kashyap 1980]). We can see from (C.5.13) that the behavior of AIC with respect to the probability of correct detection is not entirely satisfactory. Interestingly, it is precisely this kind of behavior that appears to make AIC perform satisfactorily with respect to the other possible type of performance measure, as explained below.

An alternative way of measuring the performance is to consider a more practical scenario, in which the data-generating mechanism is more complex than any of the models under test, as is usually the case in practical applications. In such a case, we can use analytical or numerical studies to determine the performance of the model picked by the rule as an *approximation* of the data-generating mechanism—for instance, we can consider the average distance between the estimated and true spectral densities or the average prediction error of the model. With respect to such a performance measure, AIC performs well, partly because of its tendency to select models with relatively large orders which may be a good thing to do in a case in which the data generating mechanism is more complex than the models used to fit it.

The nonzero overfitting probability of AIC is due to the fact that the term $2n$ in (C.5.8) (which penalizes high-order models), while larger than the term $n$ that appears in the NN rule, is still too small. Extensive simulation studies (see, e.g., [Bhansali and Downham 1977]) have found empirically that the following Generalized Information Criterion (GIC)

$$\text{GIC} = -2 \ln p_n(y, \hat{\theta}^n) + \nu n \tag{C.5.14}$$

can outperform AIC with respect to various performance measures if $\nu > 2$. Specifically, depending on the considered scenario as well as the value of $N$ and the performance measure, values of $\nu$ in the interval $\nu \in [2, 6]$ have been found to give the best performance.

In the next section, we show that GIC can be obtained as a natural theoretical extension of AIC. Hence, the use of (C.5.14) with $\nu > 2$ can be motivated on formal grounds. However, the choice of a particular $\nu$ in GIC is a more difficult problem, as we will see in Section C.6,

and cannot be solved in the current KL framework. The different framework of Section C.7 appears to be necessary to arrive at a rule having the form of (C.5.14) with a specific expression for $\nu$.

We close this section with a brief discussion on another modification of the AIC rule suggested in the literature (see, for example, [HURVICH AND TSAI 1993]). As explained before, AIC is derived by maximizing an *asymptotically* unbiased estimate of the relative KL information $I$ in (C.5.3). Interestingly, for linear-regression models (given by (C.2.2) where $\mu(\gamma)$ is a linear function of $\gamma$), the following *corrected AIC rule*, $\text{AIC}_\text{c}$, can be shown to be an *exactly* unbiased estimate of $I$:

$$\text{AIC}_\text{c} = -2 \ln p_n(y, \hat{\theta}^n) + \frac{2N}{N - n - 1} n \tag{C.5.15}$$

(See, for example, [HURVICH AND TSAI 1993; CAVANAUGH 1997].) As $N \to \infty$, $\text{AIC}_\text{c} \to \text{AIC}$ (as expected). However, for finite values of $N$, the penalty term of $\text{AIC}_\text{c}$ is larger than that of AIC. Consequently, in finite samples, $\text{AIC}_\text{c}$ has a smaller risk of overfitting than AIC, and therefore we can say that $\text{AIC}_\text{c}$ trades off a decrease of the risk of overfitting (which is rather large for AIC) for an increase in the risk of underfitting (which is quite small for AIC and hence can be slightly increased without a significant deterioration of performance). With this fact in mind, $\text{AIC}_\text{c}$ can be used as an order-selection rule for models more general than just linear regressions, even though its motivation in the general case is pragmatic rather than theoretical. For other finite-sample corrections of AIC, we refer the reader to [DE WAELE AND BROERSEN 2003; BROERSEN 2000; BROERSEN 2002; SEGHOUANE, BEKARA, AND FLEURY 2003].

## C.6 GENERALIZED CROSS-VALIDATORY KL APPROACH: THE GIC RULE

In the cross-validatory approach of the previous section, the estimation sample $x$ has the same length as the validation sample $y$. In that approach, $\hat{\theta}_x$ (obtained from $x$) is used to approximate the likelihood of the model via $E_x\{p(y, \hat{\theta}_x)\}$. The AIC rule so obtained has a nonzero probability of overfitting (even asymptotically). Intuitively, the risk of overfitting will decrease if we let the length of the validation sample be (much) larger than that of the estimation sample—that is,

$$N \triangleq \text{length}(y) = \rho \cdot \text{length}(x), \qquad \rho \geq 1 \tag{C.6.1}$$

Indeed, overfitting occurs when the model corresponding to $\hat{\theta}_x$ also fits the "noise" in the sample $x$, so that $p(x, \hat{\theta}_x)$ has a "much" larger value than the true pdf, $p(x, \theta)$. Such a model could behave reasonably well on a short validation sample $y$, but not on a long validation sample. (In the latter case, $p(y, \hat{\theta}_x)$ will take on very small values.) The simple idea in (C.6.1) of letting the lengths of the validation and estimation samples be different leads to a natural extension of AIC, as shown next.

A straightforward calculation shows that, under (C.6.1), we have

$$J_y = \rho J_x \tag{C.6.2}$$

(See, e.g., (C.2.19).) With this small difference, the calculations in the previous section carry over to the present case, and we obtain (see (C.5.4)–(C.5.5))

$$\begin{aligned} I &\simeq E_y\left\{\ln p_n(y,\hat{\theta}_y)\right\} \\ &\quad -\frac{1}{2}E_y\left\{E_x\left\{\mathrm{tr}\left(J_y\left[(\hat{\theta}_x-\theta)-(\hat{\theta}_y-\theta)\right]\left[(\hat{\theta}_x-\theta)-(\hat{\theta}_y-\theta)\right]^T\right)\right\}\right\} \\ &= E_y\left\{\ln p_n(y,\hat{\theta}_y)-\frac{1}{2}\,\mathrm{tr}\left[J_y\left(\rho J_y^{-1}+J_y^{-1}\right)\right]\right\} \\ &= E_y\left\{\ln p_n(y,\hat{\theta}_y)-\frac{1+\rho}{2}n\right\} \end{aligned} \tag{C.6.3}$$

An unbiased estimate of the right side in (C.6.3) is given by

$$\ln p(y,\hat{\theta}_y)-\frac{1+\rho}{2}n \tag{C.6.4}$$

The *generalized information criterion (GIC) rule* maximizes (C.6.4) or, equivalently, *minimizes*

$$\boxed{\mathrm{GIC}=-2\ln p_n(y,\hat{\theta}^n)+(1+\rho)n} \tag{C.6.5}$$

As expected, (C.6.5) reduces to AIC for $\rho=1$. Note also that, for a given $y$, the order selected by (C.6.5) with $\rho>1$ is always smaller than the order selected by AIC (because the penalty term in (C.6.5) is larger than that in (C.5.8)); hence, as predicted by the previous intuitive discussion, the risk of overfitting associated with GIC is smaller than for AIC when $\rho>1$.

On the negative side, there is no clear guideline for choosing $\rho$ in (C.6.5). The "optimal" value of $\rho$ in the GIC rule has been shown empirically to depend on the performance measure, the number of data samples, and the data-generating mechanism itself [McQuarrie and Tsai 1998; Bhansali and Downham 1977]. Consequently, $\rho$ should be chosen as a function of all these factors, but there is no clear rule as to how that should be done. The approach of the next section appears to be more successful than the present approach in suggesting a specific choice for $\rho$ in (C.6.5). Indeed, as we will see, that approach leads to an order-selection rule of the GIC type but with a concrete expression for $\rho$ as a function of $N$.

## C.7 BAYESIAN APPROACH: THE BIC RULE

The order-selection rule to be presented in this section can be obtained in two ways. First, let us consider *the KL framework* of the previous sections. Therefore, our goal is to maximize the relative KL information (see (C.5.1)):

$$I(p_0,\hat{p})=E_0\left\{\ln\hat{p}(y)\right\} \tag{C.7.1}$$

The ideal choice of $\hat{p}(y)$ would be $\hat{p}(y) = p_n(y, \theta^n)$. However, this choice is not possible, because the likelihood of the model, $p_n(y, \theta^n)$, is not available. Hence, we have to use a "surrogate likelihood" in lieu of $p_n(y, \theta^n)$. Let us assume, as before, that a fictitious sample $x$ is used to make inferences about $\theta$. The pdf of the estimate, $\hat{\theta}_x$, obtained from $x$ can alternatively be viewed as an *a priori* pdf of $\theta$; hence, it will be denoted by $p(\theta)$ in what follows (once again, we omit the superindex $n$ of $\theta$, $\hat{\theta}$, *etc.* to simplify the notation, whenever there is no risk for confusion). Note that we do *not* constrain $p(\theta)$ to be Gaussian. We only assume that

$$p(\theta) \text{ is flat around } \hat{\theta} \tag{C.7.2}$$

where, as before, $\hat{\theta}$ denotes the ML estimate of the parameter vector obtained from the available data sample, $y$. Furthermore, now we assume that the length of the fictitious sample is a constant that does not depend on $N$; hence,

$$p(\theta) \text{ is independent of } N \tag{C.7.3}$$

As a consequence of assumption (C.7.3), the ratio between the lengths of the validation sample and the (fictitious) estimation sample grows without bound as $N$ increases. According to the discussion in the previous section, this fact should lead to an order-selection rule having a penalty term asymptotically much larger than that of AIC or GIC (with $\rho =$ constant) and, hence, having a reduced risk of overfitting.

The scenario just introduced leads naturally to the following choice of surrogate likelihood:

$$\hat{p}(y) = E_\theta \{p(y, \theta)\} = \int p(y, \theta) p(\theta)\, d\theta \tag{C.7.4}$$

**Remark:** In the previous sections, we used a surrogate likelihood given (see (C.5.2)) by

$$\ln \hat{p}(y) = E_x \left\{ \ln p(y, \hat{\theta}_x) \right\} \tag{C.7.5}$$

However, we could have instead used a $\hat{p}(y)$ given by

$$\hat{p}(y) = E_{\hat{\theta}_x} \left\{ p(y, \hat{\theta}_x) \right\} \tag{C.7.6}$$

The rule that would be obtained by using (C.7.6) can be shown to have the same form as AIC and GIC, but with a (slightly) different penalty term. Note that the choice of $\hat{p}(y)$ in (C.7.6) is similar to the choice in (C.7.4), with the difference that for (C.7.6) the "*a priori*" pdf, $p(\hat{\theta}_x)$, depends on $N$. ■

To obtain a simple asymptotic approximation of the integral in (C.7.4), we make use of the asymptotic approximation of $p(y, \theta)$ given by (C.4.6)–(C.4.7):

$$p(y, \theta) \simeq p(y, \hat{\theta}) e^{-\frac{1}{2}(\hat{\theta}-\theta)^T \hat{J} (\hat{\theta}-\theta)} \tag{C.7.7}$$

This equation holds for $\theta$ in the vicinity of $\hat{\theta}$. Inserting (C.7.7) in (C.7.4) and using the assumption in (C.2.2), along with the fact that $p(y,\theta)$ is asymptotically much larger at $\theta=\hat{\theta}$ than at any $\theta\neq\hat{\theta}$, we obtain

$$\begin{aligned}\hat{p}(y) &\simeq p(y,\hat{\theta})p(\hat{\theta})\int e^{-\frac{1}{2}(\hat{\theta}-\theta)^T\hat{J}(\hat{\theta}-\theta)}\,d\theta\\ &= \frac{p(y,\hat{\theta})p(\hat{\theta})(2\pi)^{n/2}}{|\hat{J}|^{1/2}}\underbrace{\int\frac{1}{(2\pi)^{n/2}|\hat{J}^{-1}|^{1/2}}e^{-\frac{1}{2}(\hat{\theta}-\theta)^T\hat{J}(\hat{\theta}-\theta)}\,d\theta}_{=1}\\ &= \frac{p(y,\hat{\theta})p(\hat{\theta})(2\pi)^{n/2}}{|\hat{J}|^{1/2}}\end{aligned} \tag{C.7.8}$$

(See [Djurić 1998] and references therein for the exact conditions under which this approximation holds true.) It follows from (C.7.1) and (C.7.8) that

$$\hat{I}=\ln p(y,\hat{\theta})+\ln p(\hat{\theta})+\frac{n}{2}\ln 2\pi-\frac{1}{2}\ln|\hat{J}| \tag{C.7.9}$$

is an asymptotically unbiased estimate of the relative KL information. Note, however, that (C.7.9) depends on the *a priori* pdf of $\theta$, which has not been specified. To eliminate this dependence, we use the fact that $|\hat{J}|$ increases without bound as $N$ increases. Specifically, in most cases (but not in all; see below) we have (*cf.* (C.2.21)) that

$$\ln|\hat{J}|=\ln\left|N\cdot\frac{1}{N}\hat{J}\right|=n\ln N+\ln\left|\frac{1}{N}\hat{J}\right|=n\ln N+\mathcal{O}(1) \tag{C.7.10}$$

where we used the fact that $|cJ|=c^n|J|$ for a scalar $c$ and an $n\times n$ matrix $J$. Using (C.7.10) and the fact that $p(\theta)$ is independent of $N$ (see (C.7.3)) yields the following asymptotic approximation of the right side in (C.7.9):

$$\hat{I}\simeq\ln p_n(y,\hat{\theta}^n)-\frac{n}{2}\ln N \tag{C.7.11}$$

The *Bayesian information criterion (BIC) rule* selects the order that maximizes (C.7.11), or, equivalently, *minimizes*

$$\boxed{\text{BIC}=-2\ln p_n(y,\hat{\theta}^n)+n\ln N} \tag{C.7.12}$$

We remind the reader that (C.7.12) has been derived under the assumption that (C.2.21) holds, but this assumption is *not* always true. As an example (see [Djurić 1998] for more examples), consider once again the sinusoidal signal model with $n_c$ components (as also considered in Section C.5),

in the case of which we have (*cf.* (C.2.22)–(C.2.23)) that

$$\begin{aligned}\ln|\hat{J}| &= \ln\left|K_N^{-2}\right| + \ln\left|K_N\hat{J}K_N\right| \\ &= (2n_c+1)\ln N_s + 3n_c\ln N_s + \mathcal{O}(1) \\ &= (5n_c+1)\ln N_s + \mathcal{O}(1)\end{aligned} \tag{C.7.13}$$

Hence, in the case of *sinusoidal signals*, BIC takes on the form

$$\begin{aligned}\text{BIC} &= -2\ln p_{n_c}(y,\hat{\theta}^{n_c}) + (5n_c+1)\ln N_s \\ &= 2N_s\ln\hat{\sigma}_{n_c}^2 + (5n_c+1)\ln N_s\end{aligned} \tag{C.7.14}$$

where $\hat{\sigma}_{n_c}^2$ is as defined in (C.5.10) and $N_s$ denotes the number of complex-valued data samples.

The attribute Bayesian in the name of the rule in (C.7.12) or (C.7.14) is motivated by the use of the *a priori* pdf, $p(\theta)$, in the rule derivation, a method typical of a Bayesian approach. In fact, the BIC rule can be obtained by using a full Bayesian approach, as explained next.

To obtain the BIC rule in a *Bayesian framework*, we assume that the parameter vector $\theta$ is a random variable with a given *a priori* pdf denoted by $p(\theta)$. Owing to this assumption on $\theta$, we need to modify the previously used notation as follows: $p(y,\theta)$ will now denote the joint pdf of $y$ and $\theta$, and $p(y|\theta)$ will denote the conditional pdf of $y$ given $\theta$. Using this notation and Bayes' rule, we can write

$$p(y|H_n) = \int p_n(y,\theta^n)\,d\theta^n = \int p_n(y|\theta^n)p_n(\theta^n)\,d\theta^n \tag{C.7.15}$$

The right side of (C.7.15) is identical to that of (C.7.4). It follows from this observation and from the analysis conducted in the first part of this section that, under the assumptions (C.7.2) and (C.7.3), and asymptotically in $N$,

$$\ln p(y|H_n) \simeq \ln p_n(y,\hat{\theta}^n) - \frac{n}{2}\ln N = -\frac{1}{2}\text{BIC} \tag{C.7.16}$$

Hence, maximizing $p(y|H_n)$ is asymptotically equivalent to minimizing BIC, independently of the prior $p(\theta)$ (as long as it satisfies (C.7.2) and (C.7.3)). The rediscovery of BIC in this Bayesian framework is important: It reveals the interesting fact that the BIC rule is asymptotically equivalent to the optimal MAP rule (see Section C.3.1) and, hence, that *the BIC rule can be expected to maximize the total probability of correct detection*, at least for sufficiently large values of $N$.

The BIC rule has been proposed in [Schwarz 1978a; Kashyap 1982], among others. In [Rissanen 1978; Rissanen 1982] the same type of rule has been obtained by a different approach, one based on coding arguments and on the minimum description length (MDL) principle. The fact that the BIC rule can be derived in several different ways suggests that it might have a fundamental character. In particular, it can be shown that, under the assumption that the data-generating mechanism belongs to the model class considered, *the BIC rule is consistent*—that is,

$$\text{For BIC: the probability of correct detection} \to 1 \text{ as } N \to \infty \tag{C.7.17}$$

(See, e.g., [Söderström and Stoica 1989; McQuarrie and Tsai 1998].) This should be contrasted with the nonzero overfitting probability of AIC and GIC (with $\rho = \text{constant}$); see (C.5.12)–(C.5.13). Note that the result in (C.7.17) is not surprising in view of the asymptotic equivalence between the BIC rule and the optimal MAP rule.

Finally, we note in passing that, if we remove the condition in (C.7.3) that $p(\theta)$ be independent of $N$, then the term $\ln p(\hat{\theta})$ no longer may be eliminated from (C.7.9) by letting $N \to \infty$. Consequently, (C.7.9) would lead to a prior-dependent rule, which could be used to obtain any other rule described in this appendix by suitably choosing the prior. This line of argument can serve the theoretical purpose of interpreting various order-selection rules in a common Bayesian framework, but it appears to have little practical value; in particular, it can hardly be used to derive sound new order-selection rules.

## C.8 SUMMARY AND THE MULTIMODEL APPROACH

In the first part of this section, we summarize the model order selection rules presented in the previous sections. Then we briefly discuss and motivate the multimodel approach which, as the name suggests, is based on the idea of using more than just one model for making inferences about the signal under study.

### C.8.1 Summary

We begin with the observation that all the order-selection rules discussed in this appendix have the common form

$$-2 \ln p_n(y, \hat{\theta}^n) + \eta(n, N)n \tag{C.8.1}$$

but different *penalty coefficients* $\eta(n, N)$:

$$\begin{aligned} \text{AIC}: &\quad \eta(n, N) = 2 \\ \text{AIC}_\text{c}: &\quad \eta(n, N) = 2\frac{N}{N - n - 1} \\ \text{GIC}: &\quad \eta(n, N) = \nu = \rho + 1 \\ \text{BIC}: &\quad \eta(n, N) = \ln N \end{aligned} \tag{C.8.2}$$

Before using any of these rules for order selection in a specific problem, we need to carry out the following steps:

(i) Obtain an explicit expression for the term $-2 \ln p_n(y, \hat{\theta}^n)$ in (C.8.1). This requires the specification both of the model structures to be tested and of their postulated likelihoods. An aspect that should receive some attention here is the fact that the derivation of all previous rules assumed real-valued data and parameters. Consequently, complex-valued data and parameters must be converted to real-valued quantities in order to apply the results in this appendix.

(ii) Count the number of unknown (real-valued) parameters in each model structure under consideration. This is easily done in the parametric spectral analysis problems in which we are interested.

(iii) Verify that the assumptions that have been made to derive the rules hold true. Fortunately, most of the assumptions made are quite weak; hence, they will usually hold. Indeed, the models under test may be either nested or non-nested, and they may even be only approximate descriptions of the data-generating mechanism. However, there are two particular assumptions, made on the information matrix $J$, that do not always hold and hence must be checked. First, we assumed in all derivations that the inverse matrix, $J^{-1}$, exists; such is not always the case. Second, we made the assumption that $J$ is such that $J/N = \mathcal{O}(1)$. For some models, this is not true; when it is not true, a different normalization of $J$ is required to make it tend to a constant matrix as $N \to \infty$. (This aspect is important for the BIC rule only.)

We have used the sinusoidal signal model as an example throughout this appendix to illustrate these steps and the involved aspects.

Once these aspects have been carefully considered, we can go on to use one of the four rules in (C.8.1)–(C.8.2) for selecting the order in our estimation problem. The question of which rule should be used is not an easy one. In general, we can prefer $\text{AIC}_\text{c}$ over AIC: indeed, there is empirical evidence that $\text{AIC}_\text{c}$ outperforms AIC in small samples (whereas in medium or large samples the two rules are almost equivalent). We also tend to prefer BIC over AIC or $\text{AIC}_\text{c}$, on the grounds that BIC is an asymptotic approximation of the optimal MAP rule. Regarding GIC, as mentioned in Sections C.5 and C.6, GIC with $\nu \in [2, 6]$ (depending on the scenario under study) can outperform AIC and $\text{AIC}_\text{c}$. Hence, for lack of a more precise guideline, we can think of using GIC with $\nu = 4$, the value in the middle of the above interval. To summarize, then, a possible ranking of the four rules discussed in this appendix is as follows (the first being considered the best):

- BIC
- GIC with $\nu = 4$ ($\rho = 3$)
- $\text{AIC}_\text{c}$
- AIC

In Figure C.1, we show the penalty coefficients of the above rules, as functions of $N$, to further illustrate the relationship between them.

### C.8.2 The Multimodel Approach

We close this section with a brief discussion of a multimodel approach. Assume that we have used our favorite information criterion—call it XIC—and have computed its values for the model orders under test:

$$\text{XIC}(n); \quad n = 1, \ldots, \bar{n} \tag{C.8.3}$$

We can then pick the order that minimizes $\text{XIC}(n)$ and hence end up using a single model; this is the single-model approach.

Alternatively, we can consider a *multimodel approach*. Specifically, let us pick a $M \in [1, \bar{n}]$ (such as $M = 3$) and consider the model orders that give the $M$ smallest values of $\text{XIC}(n)$, let

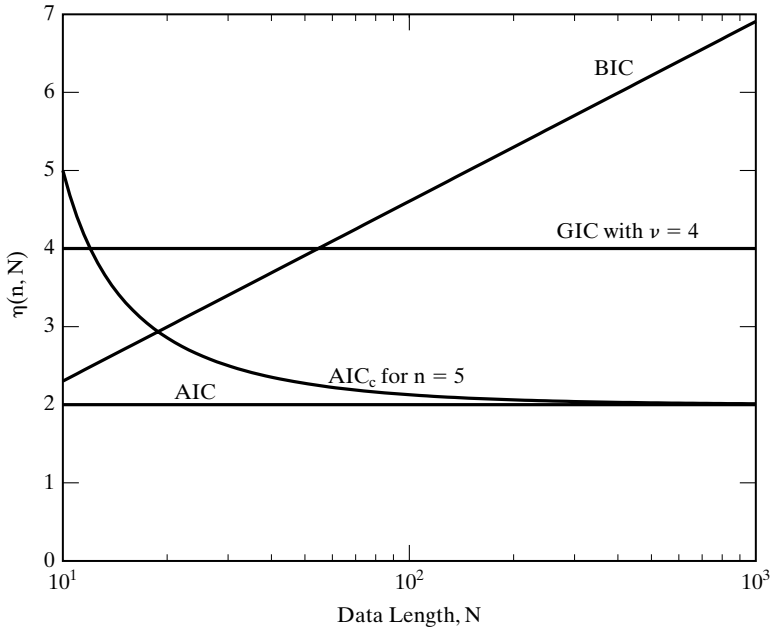


**Figure C.1** Penalty coefficients of AIC, GIC with $\nu = 4$ ($\rho = 3$), $\text{AIC}_\text{c}$ (for $n = 5$), and BIC, as functions of data length $N$.

us say $n_1, \ldots, n_M$. From the derivations presented in the previous sections of this appendix, we can see that all information criteria attempt to estimate twice the negative log-likelihood of the model:

$$-2 \ln p_n(y, \theta^n) = -2 \ln p(y|H_n) \tag{C.8.4}$$

Hence, we can use

$$e^{-\frac{1}{2}\text{XIC}(n)} \tag{C.8.5}$$

as an estimate of the likelihood of the model with order equal to $n$ (to within a multiplicative constant). Consequently, instead of using just one model corresponding to the order that minimizes $\text{XIC}(n)$, we can think of considering a combined use of the selected models (with orders $n_1, \ldots, n_M$) in which the contribution of each model is proportional to its likelihood value:

$$\frac{e^{-\frac{1}{2}\text{XIC}(n_k)}}{\sum_{j=1}^{M} e^{-\frac{1}{2}\text{XIC}(n_j)}}, \qquad k = 1, \ldots, M \tag{C.8.6}$$

For more details on the multimodel approach, including guidelines for choosing $M$, we refer the interested reader to [Burnham and Anderson 2002; Stoica, Selén, and Li 2004].