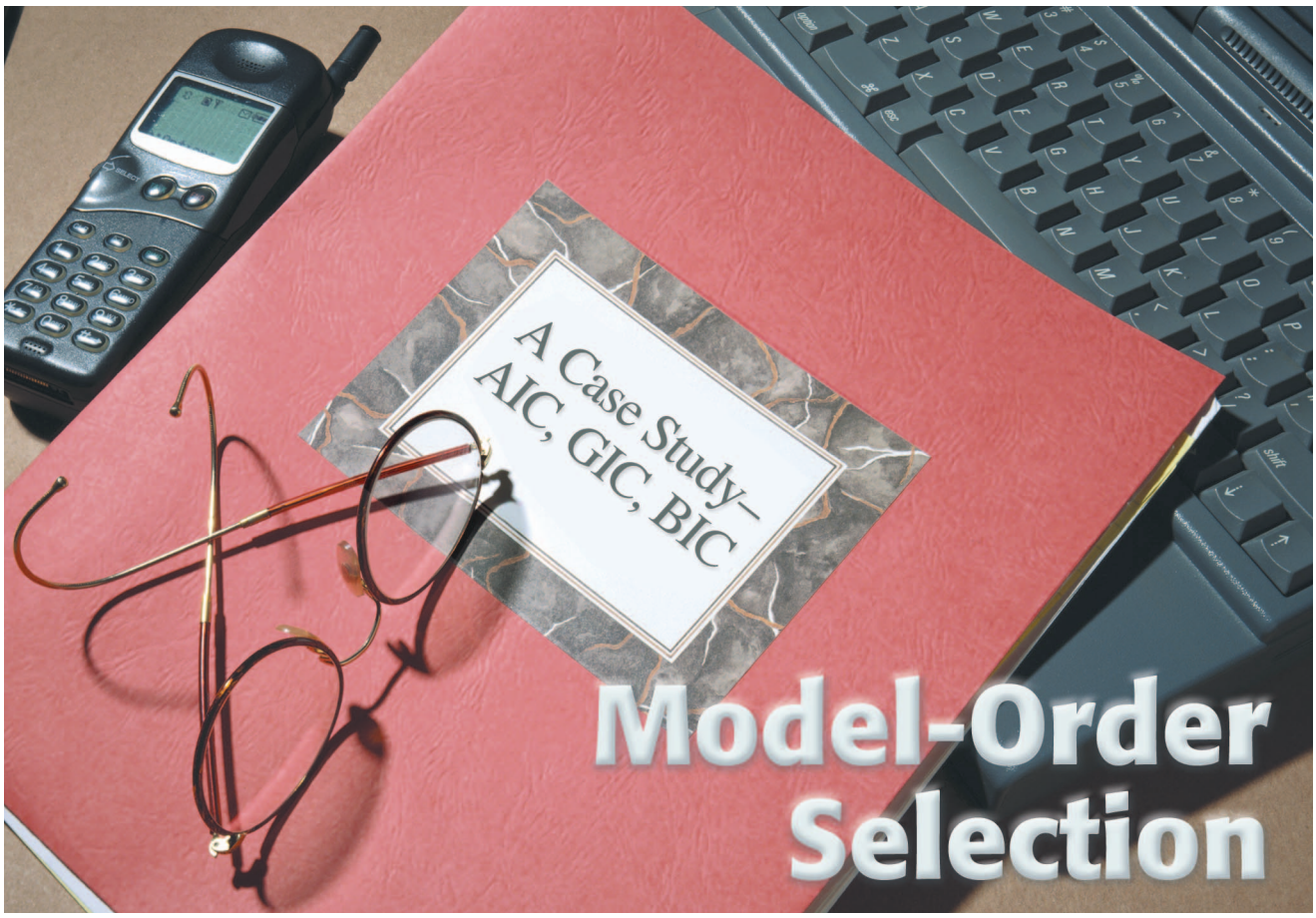


A review of information criterion rules

Petre Stoica and Yngve Selén

The parametric (or model-based) methods of signal processing often 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 a data model. Examples of these integer-valued parameters of the model include the orders of an autoregressive moving average model, the number of sinusoidal components in a sinusoids-in-noise signal, and the number of source signals impinging on a sensor array. In each of these cases, the integer-valued parameters determine the dimension of the parameter vector of the data model, and they must be estimated from the data.

In what follows we will use the following symbols:

y = the vector of available data (of size N)
θ = the (real-valued) parameter vector
n = the dimension of θ.

For short, we will refer to n as the model order, even though sometimes n is not really an order (see, e.g., the above examples). We assume that both y and θ are real valued

$$y \in \mathcal{R}^N$$
$$\theta \in \mathcal{R}^n.$$

Whenever we need to emphasize that the number of elements in θ is n, we will use the notation θ^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 signal processing. The nonparametric methods do not have such a requirement.

The literature on order selection is as considerable as that on (real-valued) parameter estimation (see, e.g., [1]–[7] and the references therein). However, many order-selection rules are tied to specific parameter estimation methods, and, 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. The MLM is

order selection rule associated with MAP?

1053-5888/04/$20.00©2004IEEE

likely the most commonly used parameter estimation method. Consequently, the order-estimation rules that can be used with the MLM are of quite general interest.

Maximum Likelihood Parameter Estimation

In this section, we review briefly the MLM of parameter estimation and some of its main properties that are of interest in this article. Let

is this class conditional PDF? I feel so

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

The maximum likelihood (ML) estimate of θ, which we denote by $\hat{\theta}$, is given by the maximizer of $p(y, \theta)$; see, e.g., [2], [8]–[15]. Alternatively, as $\ln(\cdot)$ is a monotonically increasing function

$$\hat{\theta} = \arg \max_{\theta} \ln\, p(y, \theta). \quad (1)$$

relate this to MLE seen before

Under the Gaussian data assumption, the MLM typically reduces to the nonlinear least-squares (NLS) method of parameter estimation. To illustrate this fact, let us assume that the observation vector y can be written as

$$y = \mu(\gamma) + e \quad (2)$$

Remember Get math intuition for this

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

$$p(y, \theta) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} e^{-\frac{||y-\mu(\gamma)||^2}{2\sigma^2}} \quad (3)$$

where

$$\theta = \begin{bmatrix} \gamma \\ \sigma^2 \end{bmatrix}. \quad (4)$$

We deduce from (3) that

$$-2 \ln\, p(y, \theta) = N \ln 2\pi + N \ln \sigma^2 + \frac{||y - \mu(\gamma)||^2}{\sigma^2}. \quad (5)$$

A simple calculation based on (5) shows that the ML estimates of γ and σ^2 are given by

$$\hat{\gamma} = \arg \min_{\gamma} ||y - \mu(\gamma)||^2 \quad (6)$$

$$\hat{\sigma}^2 = \frac{1}{N} ||y - \mu(\hat{\gamma})||^2. \quad (7)$$

The corresponding value of the likelihood function is given by

$$-2 \ln\, p\big(y, \hat{\theta}\big) = \text{constant} + N \ln \hat{\sigma}^2. \quad (8)$$

As can be seen from (6), in the present case the MLM indeed reduces to the NLS.

A special case of (2), which we will address in this article, is the sinusoidal signal model

$$y_c(t) = \sum_{k=1}^{n_c} \alpha_k e^{i(\omega_k t + \varphi_k)} + e(t), \quad t = 1, \ldots, N_s \quad (9)$$

where $\{[\alpha_k\ \omega_k\ \varphi_k]\}$ denote the amplitude, frequency, and phase of the kth sinusoidal component; N_s is the number of observed complex-valued samples; n_c is the number of sinusoidal components present in the signal; and $e(t)$ is the observation noise. In this case

$$N = 2N_s \quad (10)$$

$$n = 3n_c + 1. \quad (11)$$

We will use the sinusoidal signal model in (9) as a vehicle for illustrating how the various general order-selection rules presented in what follows should be used in a specific situation.

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 θ and covariance equal to the Cramér-Rao bound (CRB) matrix (see, e.g., [2], [16] for a discussion about the CRB). Consequently, asymptotically in N, the PDF of $\hat{\theta}$ is given by

$$p\big(\hat{\theta}\big) = \frac{1}{(2\pi)^{n/2} |J^{-1}|^{1/2}} e^{-\frac{1}{2}\left(\hat{\theta}-\theta\right)^T J \left(\hat{\theta}-\theta\right)} \quad (12)$$

where

$$J = -E\left\{\frac{\partial^2 \ln p(y, \theta)}{\partial\theta\ \partial\theta^T}\right\} \quad (13)$$

is the so-called (Fisher) information matrix.

Note: To simplify the notation, we use the symbol θ for both the true parameter vector and the parameter vector viewed as an unknown variable. The exact meaning of θ should be clear from the context.

The "regularity conditions" referred to above require that n is not 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 most parametric signal processing problems but not for all (see, e.g., [17]–[19]).

To close this section, we note that under mild conditions

$$\left[-\frac{1}{N}\frac{\partial^2 \ln p(y, \theta)}{\partial\theta\ \partial\theta^T} - \frac{1}{N} J\right] \to 0 \quad \text{as } N \to \infty. \quad (14)$$

To motivate (14) for the fairly general data model in (2), we can argue as follows. Let us rewrite the negative log-likelihood function associated with (2) as [see (5)]

$$-\ln\, p(y,\theta) = \text{constant} + \frac{N}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]^2 \tag{15}$$

where the subindex t denotes the tth component. From (15) we obtain by a simple calculation:

$$-\frac{\partial \ln\, p(y,\theta)}{\partial\theta} = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{t=1}^{N}[y_t - \mu_t(\gamma)]\mu_t'(\gamma) \\ \frac{N}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{t=1}^{N}[y_t - \mu_t(\gamma)]^2 \end{bmatrix} \tag{16}$$

where

$$\mu_t'(\gamma) = \frac{\partial \mu_t(\gamma)}{\partial \gamma}. \tag{17}$$

Differentiating (16) once again gives (18) (shown at the bottom of the page) where $e_t = y_t - \mu_t(\gamma)$ and

$$\mu_t''(\gamma) = \frac{\partial^2 \mu_t(\gamma)}{\partial\gamma\ \partial\gamma^T}. \tag{19}$$

Taking the expectation of (18) and dividing by N, we get

$$\frac{1}{N}J = \begin{bmatrix} \frac{1}{\sigma^2}\left(\frac{1}{N}\sum_{t=1}^{N}\mu_t'(\gamma)\mu_t'^T(\gamma)\right) & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}. \tag{20}$$

We assume that $\mu(\gamma)$ is such that the above matrix has a finite limit as $N \to \infty$. Under this assumption, and the previously made assumption on e, we can also show from (18) 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 (20), which concludes the motivation of (14). Letting

$$\hat{J} = -\left.\frac{\partial^2 \ln\, p(y,\,\theta)}{\partial\theta\ \partial\theta^T}\right|_{\theta=\hat{\theta}} \tag{21}$$

we deduce from (14), (20), and the consistency of $\hat{\theta}$ that, for sufficiently large values of N,

$$\frac{1}{N}\hat{J} \approx \frac{1}{N}J = \mathcal{O}(1). \tag{22}$$

Hereafter, $\approx$ 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 (20) has a finite limit, as $N \to \infty$, holds for many problems but *not* for the sinusoidal parameter estimation problem associated with (9). In the latter case, (22) needs to be modified as (see, e.g., [20] and [21])

$$K_N\hat{J}K_N \approx K_N J K_N = \mathcal{O}(1) \tag{23}$$

where

$$K_N = \begin{bmatrix} \frac{1}{N_s^{3/2}}I_{n_c} & 0 \\ 0 & \frac{1}{N_s^{1/2}}I_{2n_c+1} \end{bmatrix} \tag{24}$$

and where I_k denotes the $k \times k$ identity matrix. To write (24), 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 article, as we will see later on.

Useful Mathematical Preliminaries and Outlook

In this section, we discuss a number of mathematical tools that will be used in the following 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. We first 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 certain sense). Then we discuss the Kullback-Leibler (KL) information criterion, which lies at the basis of another approach that can be used to derive model order-selection rules.

MAP Selection Rule

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

$$n \in [1, \overline{n}]. \tag{25}$$

We assume that the hypotheses $\{H_n\}_{n=1}^{\overline{n}}$ are mutually exclusive (i.e., only one of them can hold true at a time). As an example, for a real-valued autoregres-

$$-\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\ \partial\theta^T} = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{t=1}^{N} e_t\mu_t''(\gamma) + \frac{1}{\sigma^2}\sum_{t=1}^{N}\mu_t'(\gamma)\mu_t'^T(\gamma) & \frac{1}{\sigma^4}\sum_{t=1}^{N} e_t\mu_t'(\gamma) \\ \frac{1}{\sigma^4}\sum_{t=1}^{N} e_t\mu_t'(\gamma) & -\frac{N}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{t=1}^{N} e_t^2 \end{bmatrix} \tag{18}$$

sive (AR) signal with coefficients $\{a_k\}$ we can define H_n as

$$H_n: \ a_n \neq 0 \text{ and } a_{n+1} = \cdots = a_{\overline{n}} = 0. \tag{26}$$

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 (11)

$$n = 3n_c + 1. \tag{27}$$

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

$$\begin{aligned} H_{n_c}&: \alpha_k \neq 0 \text{ for } k = 1, \ldots, n_c \text{ and} \\ &\alpha_k = 0 \text{ for } k = n_c + 1, \ldots, \bar{n}_c \end{aligned} \tag{28}$$

for $n_c \in [1, \overline{n}_c]$ [with the corresponding "model order," n, being given by (27)].

Note: The hypotheses $\{H_n\}$ can be either nested or nonnested. 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 above H_1 and

H_3: the signal consists of one sinusoid in noise

are nonnested.

Let

$$p_n(y|H_n) = \text{the PDF of } y \text{ under } H_n. \tag{29}$$

Whenever we want to emphasize the possible dependence of the PDF in (29) on the parameter vector of the model corresponding to H_n, we write

$$p_n(y, \theta^n) \stackrel{\triangle}{=} p_n(y|H_n). \tag{30}$$

Assuming that (29) 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

$$p_n(H_n|y) = \frac{p_n(y|H_n)p_n(H_n)}{p(y)}. \tag{31}$$

The MAP probability rule selects the order n (or the hypothesis H_n) that maximizes (31). As the denominator in (31) does not depend on n, the MAP rule is given by

$$\max_{n \in [1, \overline{n}]} p_n(y|H_n)p_n(H_n). \tag{32}$$

Most typically, the hypotheses $\{H_n\}$ are a priori equiprobable, i.e.,

$$p_n(H_n) = \frac{1}{\overline{n}}, \quad n = 1, \ldots, \overline{n} \tag{33}$$

in which case the MAP rule reduces to [see (32)]

$$\max_{n \in [1, \overline{n}]} p_n(y|H_n). \tag{34}$$

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

$$\begin{aligned} P_{cd} = \Pr\{&[(\text{decide } H_1) \cap (H_1 = \text{true})] \cup \cdots \cup \\ &\times [(\text{decide } H_{\overline{n}}) \cap (H_{\overline{n}} = \text{true})]\}. \end{aligned} \tag{35}$$

The attribute "average" that has been attached to P_{cd} is motivated by the fact that (35) 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 was two (let us say), which is $\Pr\{\text{decide } H_2 | H_2\}$).

Note: 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, $\mathcal{R}^N$, into $\overline{n}$ sets $\{S_n\}_{n=1}^{\overline{n}}$, which are such that

$$\text{We decide } H_n \text{ if and only if } y \in S_n. \tag{36}$$

Making use of (36) along with the fact that the hypotheses $\{H_n\}$ are mutually exclusive, we can write P_{cd} in (35) as

$$\begin{aligned} P_{cd} &= \sum_{n=1}^{\overline{n}} \Pr\{(\text{decide } H_n) \cap (H_n = \text{true})\} \\ &= \sum_{n=1}^{\overline{n}} \Pr\{(\text{decide } H_n) \mid H_n\} \Pr\{H_n\} \\ &= \sum_{n=1}^{\overline{n}} \int_{S_n} p_n(y|H_n)p_n(H_n)\, dy \\ &= \int_{\mathcal{R}^N} \left[\sum_{n=1}^{\overline{n}} I_n(y)p_n(y|H_n)p_n(H_n) \right] dy \end{aligned} \tag{37}$$

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

$$I_n(y) = \begin{cases} 1, & \text{if } y \in S_n \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

Next, observe that for any given data vector y one and only one indicator function can be equal to one (as the sets S_n do not overlap and their union is $\mathcal{R}^N$). This observation, along with (37) for P_{cd}, implies that the MAP rule in (32) maximizes P_{cd}, as stated. Note that the sets $\{S_n\}$ corresponding to the MAP rule are implicitly

defined via (32); however, $\{S_n\}$ were of no real interest in the proof, as both they and the indicator functions were introduced only to simplify the above proof. For more details on the topic of this subsection, see [14] and [22].

KL 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 expressed using the KL information or discrepancy function (see [23])

$$D(p_0, \hat{p}) = \int p_0(y) \ln \left[\frac{p_0(y)}{\hat{p}(y)} \right] dy. \tag{39}$$

(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 (39) as

$$\begin{aligned} D(p_0, \hat{p}) &= E_0 \left\{ \ln \left[\frac{p_0(y)}{\hat{p}(y)} \right] \right\} \\ &= E_0 \left\{ \ln p_0(y) \right\} - E_0 \left\{ \ln \hat{p}(y) \right\}. \end{aligned} \tag{40}$$

Next, we prove that (39) possesses the properties of a suitable discrepancy function

$$\begin{aligned} &D(p_0, \hat{p}) \geq 0 \\ &D(p_0, \hat{p}) = 0 \quad \text{if and only if } p_0(y) = \hat{p}(y). \end{aligned} \tag{41}$$

To verify (41), we use the fact that

$$-\ln \lambda \geq 1 - \lambda \quad \text{for any } \lambda > 0 \tag{42}$$

and

$$-\ln \lambda = 1 - \lambda \quad \text{if and only if } \lambda = 1. \tag{43}$$

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

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

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

The KL discrepancy function can be viewed as showing 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 the following three sections).

Outlook: Theoretical and Practical Perspectives

Neither the MAP rule nor the KL information can be directly used for order selection because the PDFs of the data vector under the various hypotheses or the true data PDF are usually unknown. A possible way of using the MAP approach for order detection consists of assuming an a priori PDF for the unknown parameter vector θ^n and integrating θ^n out of $p_n(y, \theta^n)$ to obtain $p_n(y|H_n)$. This Bayesian-type approach will be discussed later in the article. Regarding the KL approach, a natural way of using it for order selection consists of 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 determining the model order by minimizing $\hat{D}(p_0, \hat{p})$. This KL-based approach will be discussed in the following sections.

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. As this assumption is unlikely to hold in applications with real-life data, the reader may justifiably wonder whether an order-selection rule derived under such an assumption has any practical value. To address this concern, let us remark on the fact 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 ML, KL, and MAP) are likely to perform well in applications despite the fact that some of the assumptions made when deriving them do not hold exactly. While the precise behavior of order-selection rules (such as those presented in the sections to follow) in various mismodeling scenarios is not well understood, extensive simulation results (see, e.g., [3]–[5]) lend support to the above claim.

Direct KL Approach: No-Name Rule

The model-dependent part of the KL discrepancy (40) is given by

$$-E_0 \left\{ \ln \hat{p}(y) \right\} \tag{44}$$

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 this index later on, when needed.) Minimization of (44) with respect to the model order is equivalent to maximization of the function

$$I(p_0, \hat{p}) = E_0 \left\{ \ln \hat{p}(y) \right\}, \tag{45}$$

which is sometimes called the relative KL information. The ideal choice for $\hat{p}(y)$ in (45) 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

$$\hat{p}(y) = p\left(y, \hat{\theta}\right) \tag{46}$$

in (45), which would give

$$I\left(p_0, p\left(y, \hat{\theta}\right)\right) = E_0\left\{\ln p\left(y, \hat{\theta}\right)\right\}. \tag{47}$$

Because the true PDF of the data vector is unknown, we cannot evaluate the expectation in (47). Apparently, what we could easily do is to use the following unbiased estimate of $I(p_0, p(y, \hat{\theta}))$, instead of (47) itself,

$$\hat{I} = \ln p\left(y, \hat{\theta}\right). \tag{48}$$

The order-selection rule that maximizes (48) does *not* have satisfactory properties, however. This is especially true for nested models, in the case of which the order-selection rule based on the maximization of (48) fails completely. Indeed, for nested models this rule will always choose the maximum possible order $\overline{n}$ owing to the fact that $\ln p_n(y, \hat{\theta}^n)$ monotonically increases 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 using the so-obtained approximation to define $\ln \hat{p}(y)$ in (45)

$$\begin{aligned}\ln p_n(y, \theta^n) \approx{}& \ln p_n(y, \hat{\theta}^n) \\ &+ \left(\theta^n - \hat{\theta}^n\right)^T \left[\left.\frac{\partial \ln p_n(y, \theta^n)}{\partial \theta^n}\right|_{\theta^n = \hat{\theta}^n}\right] \\ &+ \frac{1}{2}\left(\theta^n - \hat{\theta}^n\right)^T \left[\left.\frac{\partial^2 \ln p_n(y, \theta^n)}{\partial \theta^n\, \partial \theta^{nT}}\right|_{\theta^n = \hat{\theta}^n}\right] \\ &\times \left(\theta^n - \hat{\theta}^n\right) \triangleq \ln \hat{p}_n(y). \end{aligned} \tag{49}$$

Because $\hat{\theta}^n$ is the maximizer of $\ln p_n(y, \theta^n)$, the second term in (49) is equal to zero. Hence, we can write [see also (22)]

$$\ln \hat{p}_n(y) \approx \ln p_n\left(y, \hat{\theta}^n\right) - \frac{1}{2}\left(\theta^n - \hat{\theta}^n\right)^T J\left(\theta^n - \hat{\theta}^n\right). \tag{50}$$

According to (12),

$$\begin{aligned} &E_0\left\{\left(\theta^n - \hat{\theta}^n\right)^T J\left(\theta^n - \hat{\theta}^n\right)\right\} \\ &\quad = \mathrm{tr}\left[J E_0\left\{\left(\theta^n - \hat{\theta}^n\right)\left(\theta^n - \hat{\theta}^n\right)^T\right\}\right] \\ &\quad = \mathrm{tr}[I_n] = n, \end{aligned} \tag{51}$$

which means that, for the choice of $\hat{p}_n(y)$ in (50), we have

$$I = E_0\left\{\ln p_n\left(y, \hat{\theta}^n\right) - \frac{n}{2}\right\}. \tag{52}$$

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

$$\ln p_n\left(y, \hat{\theta}^n\right) - \frac{n}{2}. \tag{53}$$

The corresponding order-selection rule maximizes (53) or, equivalently, minimizes

$$\mathrm{NN}(n) = -2 \ln p_n\left(y, \hat{\theta}^n\right) + n \tag{54}$$

with respect to the model order n. This no-name (NN) rule can be shown to perform better than that based on (48) but worse than the rules presented in the following sections. Essentially, the problem with (54) 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 (54) decreases with increasing n (for nested models), whereas the second term increases. Hence, the second term in (54) 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 (54) but with a larger penalty term, and they do have better properties than (54). Despite this fact, we have chosen to present (54) briefly in this section for two reasons: 1) 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 2) this section has laid the groundwork for the derivation of better order-selection rules based on the KL approach in the next two sections.

To close the present section, we motivate the multiplication with -2 in going from (53) to (54). The reason we prefer (54) to (53) is simply due to the fact that for the fairly common NLS model in (2) and the associated Gaussian likelihood in (3), $-2 \ln p_n(y, \hat{\theta}^n)$ takes on the following convenient form:

$$-2 \ln p_n\left(y, \hat{\theta}^n\right) = N \ln \hat{\sigma}_n^2 + \text{constant} \tag{55}$$

[see (5)–(7)]. Hence, in such a case we can replace $-2 \ln p_n(y, \hat{\theta}^n)$ in (54) 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 (54).

Cross-Validatory KL Approach: The Akaike Information Criterion 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 maximizing the relative KL information [see (45)]

$$I\left(p_0, \hat{p}\right) = E_0\left\{\ln \hat{p}(y)\right\}. \tag{56}$$

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 already explained, ideally we would like to use the true PDF of the model as $\hat{p}(y)$ in (56), i.e., $\hat{p}(y) = p_n(y, \theta^n)$, but this is not possible since $p_n(y, \theta^n)$ is unknown. Hence, we have to choose

$\hat{p}(y)$ in a different way. This choice is important, as it eventually determines the model order-selection rule that we will obtain.

The other issue we should consider when using the approach based on (56) is that the expectation in (56) 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 (56).

Let x denote a fictitious data vector with the same size N and the same PDF as y but which 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:

$$\ln \hat{p}(y) = E_x\{\ln p(y, \hat{\theta}_x)\} \tag{57}$$

which, when inserted in (56), yields

$$I = E_y\left\{E_x\{\ln p(y, \hat{\theta}_x)\}\right\}. \tag{58}$$

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 [24] and [25], has an interesting cross-validation interpretation: we use the sample x for estimation and the independent sample y for validation of the so-obtained model's PDF. Note that the dependence of (58) on the fictitious sample x is eliminated (as it should be, since 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 (49) and (50), yields

$$\begin{aligned}\ln p(y, \hat{\theta}_x) &\approx \ln p(y, \hat{\theta}_y) \\ &\quad + (\hat{\theta}_x - \hat{\theta}_y)^T \left[\left.\frac{\partial \ln p(y, \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_y}\right] \\ &\quad + \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)^T \left[\left.\frac{\partial^2 \ln p(y, \theta)}{\partial \theta\, \partial \theta^T}\right|_{\theta=\hat{\theta}_y}\right] \\ &\quad \times (\hat{\theta}_x - \hat{\theta}_y) \approx \ln p(y, \hat{\theta}_y) \\ &\quad - \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)^T J_y (\hat{\theta}_x - \hat{\theta}_y)\end{aligned} \tag{59}$$

where J_y is the J matrix, as defined in (21), associated with the data vector y. Using the fact that x and y have the same PDF (which, in particular, implies that $J_y = J_x$), along with the fact that they are independent of each other, we can show that

$$\begin{aligned}&E_y\left\{E_x\{(\hat{\theta}_x - \hat{\theta}_y)^T J_y (\hat{\theta}_x - \hat{\theta}_y)\}\right\} \\ &\quad = E_y\left\{E_x\left\{\mathrm{tr}\Big(J_y[(\hat{\theta}_x - \theta) - (\hat{\theta}_y - \theta)] \right.\right. \\ &\qquad\qquad \left.\left. \times [(\hat{\theta}_x - \theta) - (\hat{\theta}_y - \theta)]^T\Big)\right\}\right\} \\ &\quad = \mathrm{tr}\left[J_y\left(J_x^{-1} + J_y^{-1}\right)\right] = 2n.\end{aligned} \tag{60}$$

Inserting (60) in (59) yields the following asymptotic approximation of the relative KL information in (58):

$$I \approx E_y\left\{\ln p_n(y, \hat{\theta}^n) - n\right\} \tag{61}$$

(where we have omitted the subindex y of $\hat{\theta}$ but reinstated the superindex n). Evidently, (61) can be estimated in an unbiased manner by

$$\ln p_n(y, \hat{\theta}^n) - n. \tag{62}$$

Maximizing (62) with respect to n is equivalent to minimizing the following function of n

$$\mathrm{AIC} = -2 \ln p_n(y, \hat{\theta}^n) + 2n \tag{63}$$

where AIC stands for Akaike information criterion (the reasons for multiplying (62) by -2 to get (63) and for the use of the word "information" in the name given to (63) have been explained before; see the previous two sections).

As an example, for the sinusoidal signal model with n_c components [see (9)], AIC takes on the following form [see (6)–(11)]:

$$\mathrm{AIC} = 2N_s \ln \hat{\sigma}^2_{n_c} + 2(3n_c + 1) \tag{64}$$

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

$$\hat{\sigma}^2_{n_c} = \frac{1}{N_s}\sum_{t=1}^{N_s}\left|y_c(t) - \sum_{k=1}^{n_c}\hat{\alpha}_k e^{i(\hat{\omega}_k t + \hat{\varphi}_k)}\right|^2. \tag{65}$$

Note: AIC can also be obtained by using the following relative KL information function, in lieu of (58),

$$I = E_y\left\{E_x\{\ln p(x, \hat{\theta}_y)\}\right\}. \tag{66}$$

Note that, in (66), x is used for validation and y for estimation. However, the derivation of AIC from (66) is more complicated; such a derivation, which is left as an exercise to the reader, will make use of two Taylor series expansions and 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, e.g., [3]–[6]). The

performance of a model order-selection rule, such as AIC, can be measured in different ways.

As a first possibility, we can consider a scenario in which the data-generating mechanism belongs to the class of models under test and thus there is a true order. In such a case, 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

$$\text{the probability of underfitting} \to 0 \tag{67}$$

$$\text{the probability of overfitting} \to \text{constant} > 0 \tag{68}$$

as $N \to \infty$ (see, e.g., [3], [26]). We can see from (68) 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, which is usually the case in practical applications. In such a case we can use 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 [see (68)] is due to the fact that the term $2n$ in (63) (that penalizes high-order models), while larger than the term n that appears in the NN rule, is still too small. In effect, extensive simulation studies (see, e.g., [27]) have empirically found that the following generalized information criterion (GIC):

$$\text{GIC} = -2 \ln p_n\big(y, \hat{\theta}^n\big) + \nu n \tag{69}$$

may 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 ν 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 (69) with $\nu > 2$ can be motivated on formal grounds. However, the choice of ν in GIC is a more difficult problem that cannot be solved in the current KL framework (see the next section for details). The different Bayesian approach, presented later in this article, appears to be necessary to arrive at a rule having the form of (69) but with a specific expression for ν.

We close this section with a brief discussion on another modification of the AIC rule suggested in the literature (see, e.g., [28]). As explained before, AIC is derived by maximizing an asymptotically unbiased estimate of the relative KL information I in (58). Interestingly, for linear regression models (given by (2) where $\mu(\gamma)$ is a linear function of γ), the following corrected AIC rule, AIC_c, can be shown to be an exactly unbiased estimate of I

$$\text{AIC}_\text{c} = -2 \ln p_n\big(y, \hat{\theta}^n\big) + \frac{2N}{N - n - 1} n \tag{70}$$

(see, e.g., [28] and [29]). As $N \to \infty$, $\text{AIC}_\text{c} \to \text{AIC}$ (as expected). For finite values of N, however, the penalty term of AIC_c is larger than that of AIC. Consequently, in finite samples AIC_c has a smaller risk of overfitting than AIC, and therefore we can say that AIC_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 it can be slightly increased without a significant deterioration of performance). With this fact in mind, AIC_c can be used as an order-selection rule for more general models 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 [30]–[32].

Generalized Cross-Validatory KL Approach: The GIC Rule

In the cross-validatory approach of the previous section, the estimation sample x had the same length as the validation sample y. In that approach, $\hat{\theta}_x$ (obtained from x) was 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, i.e.,

$$N \triangleq \text{length}(y) = \rho \cdot \text{length}(x), \quad \rho \geq 1. \tag{71}$$

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 may 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 (71) of letting the lengths of the validation and estimation samples be different leads to a natural extension of AIC, as shown below.

A straightforward calculation shows that under (71) we have

$$J_y = \rho J_x \tag{72}$$

[see, e.g., (20)]. With this small difference, the calculations in the previous section carry over to the present case, and we obtain [see (59)–(60)]

$$\begin{aligned} I &\approx E_y\{\ln p(y,\hat{\theta}_y)\} \\ &\quad -\frac{1}{2}E_y\Big\{E_x\Big\{\mathrm{tr}\Big(J_y\big[(\hat{\theta}_x-\theta)-(\hat{\theta}_y-\theta)\big] \\ &\qquad\qquad \times\big[(\hat{\theta}_x-\theta)-(\hat{\theta}_y-\theta)\big]^T\Big)\Big\}\Big\} \\ &= E_y\left\{\ln p(y,\hat{\theta}_y)-\frac{1}{2}\mathrm{tr}\big[J_y(\rho J_y^{-1}+J_y^{-1})\big]\right\} \\ &= E_y\left\{\ln p(y,\hat{\theta}_y)-\frac{1+\rho}{2}n\right\}. \end{aligned} \tag{73}$$

An unbiased estimate of the right side in (73) is given by

$$\ln p(y,\hat{\theta}_y)-\frac{1+\rho}{2}n. \tag{74}$$

The generalized information criterion (GIC) rule maximizes (74) or, equivalently, *minimizes*

$$\mathrm{GIC}=-2\ln p_n(y,\hat{\theta}^n)+(1+\rho)n. \tag{75}$$

As expected, (75) reduces to AIC for $\rho=1$. Also note that, for a given y, the order selected by (75) with $\rho>1$ is always smaller than the order selected by AIC [because the penalty term in (75) is larger than that in (63)]; hence, as predicted by the previous intuitive discussion, the risk of overfitting associated with GIC is smaller than for AIC (for $\rho>1$).

On the negative side, there is no clear guideline for choosing ρ in (75). As already mentioned in the previous section, the "optimal" value of ρ in the GIC rule was empirically shown to depend on the performance measure, the number of data samples, and the data-generating mechanism itself [3], [27]. Consequently, ρ should be chosen as a function of all these factors but, as already stated, there is no clear hint as to how that could be done. The approach of the next section appears to be more successful than the present approach in suggesting a specific choice for ρ in (75). Indeed, as we will see, that approach leads to an order-selection rule of the GIC type but with a clear expression for ρ as a function of N.

Bayesian Approach: The Bayesian Information Criterion 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 (56)]

$$I(p_0,\hat{p})=E_0\{\ln\hat{p}(y)\}. \tag{76}$$

The ideal choice of $\hat{p}(y)$ would be $\hat{p}(y)=p_n(y,\theta^n)$. This choice is not possible, however, since 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 was used to make inferences about θ. The PDF of the estimate $\hat{\theta}_x$ obtained from x can alternatively be viewed as an a priori PDF of θ and, hence, it will be denoted by $p(\theta)$ in what follows. (Once again, we omit the superindex n of θ, $\hat{\theta}$, 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

$$p(\theta)\text{ is flat around }\hat{\theta} \tag{77}$$

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, which implies that

$$p(\theta)\text{ is independent of }N. \tag{78}$$

As a consequence of assumption (78), 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 with an asymptotically much larger penalty term than that of AIC or GIC (with $\rho=$ constant) and, hence, with a reduced risk of overfitting.

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

$$\hat{p}(y)=E_\theta\{p(y,\theta)\}=\int p(y,\theta)p(\theta)d\theta. \tag{79}$$

Note: In the previous sections we used a surrogate likelihood given by [see (57)]

$$\ln\hat{p}(y)=E_x\{\ln p(y,\hat{\theta}_x)\}. \tag{80}$$

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

$$\hat{p}(y)=E_{\hat{\theta}_x}\{p(y,\hat{\theta}_x)\}. \tag{81}$$

The rule that would be obtained by using (81) can be shown to have the same form as AIC/GIC but with a (slightly) different penalty term. Note that the choice of $\hat{p}(y)$ in (81) is similar to the choice in (79) considered in this section, with the difference that for (81) the "a priori" PDF, $p(\hat{\theta}_x)$, depends on N.

To obtain a simple asymptotic approximation of the integral in (79) we make use of the asymptotic approximation of $p(y,\theta)$ given by (49)–(50)

$$p(y,\theta)\approx p(y,\hat{\theta})e^{-\frac{1}{2}(\hat{\theta}-\theta)^T\hat{J}(\hat{\theta}-\theta)}, \tag{82}$$

which holds for θ in the vicinity of $\hat{\theta}$. Inserting (82) in

(79) and using the assumption in (77) 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

$$\begin{aligned}\hat{p}(y) &\approx p\big(y,\hat{\theta}\big)p\big(\hat{\theta}\big)\int e^{-\frac{1}{2}\left(\hat{\theta}-\theta\right)^T\hat{J}\left(\hat{\theta}-\theta\right)}\,d\theta\\ &= \frac{p\big(y,\hat{\theta}\big)p\big(\hat{\theta}\big)(2\pi)^{n/2}}{|\hat{J}|^{1/2}}\\ &\quad\times\underbrace{\int\frac{1}{(2\pi)^{n/2}|\hat{J}^{-1}|^{1/2}}e^{-\frac{1}{2}\left(\hat{\theta}-\theta\right)^T\hat{J}\left(\hat{\theta}-\theta\right)}\,d\theta}_{=1}\\ &= \frac{p\big(y,\hat{\theta}\big)p\big(\hat{\theta}\big)(2\pi)^{n/2}}{|\hat{J}|^{1/2}}\end{aligned}\tag{83}$$

(see [21] and the references therein for the exact conditions under which the above approximation holds true). It follows from (76) and (83) that

$$\hat{I}=\ln p\big(y,\hat{\theta}\big)+\ln p\big(\hat{\theta}\big)+\frac{n}{2}\ln 2\pi-\frac{1}{2}\ln|\hat{J}|\tag{84}$$

is an asymptotically unbiased estimate of the relative KL information. Note, however, that (84) depends on the a priori PDF of θ, 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 that [cf. (22)]

$$\begin{aligned}\ln|\hat{J}| &= \ln\left|N\cdot\frac{1}{N}\hat{J}\right|\\ &= n\ln N+\ln\left|\frac{1}{N}\hat{J}\right|\\ &= n\ln N+\mathcal{O}(1)\end{aligned}\tag{85}$$

where we used the fact that $|cJ|=c^n|J|$ for a scalar c and an $n\times n$ matrix J. Using (85) and the fact that $p(\theta)$ is independent of N [see (78)] yields the following asymptotic approximation of the right side in (84)

$$\hat{I}\approx\ln p_n\big(y,\hat{\theta}^n\big)-\frac{n}{2}\ln N.\tag{86}$$

The Bayesian information criterion (BIC) rule selects the order that maximizes (86) or, equivalently, minimizes

$$\mathrm{BIC}=-2\ln p_n\big(y,\hat{\theta}^n\big)+n\ln N.\tag{87}$$

We remind the reader that (87) has been derived under the assumption that (22) holds, which is *not* always true. As an example (see [21] for more examples), consider once again the sinusoidal signal model with n_c components, in the case of which we have [cf. (23) and (24)]

$$\begin{aligned}\ln|\hat{J}| &= \ln\left|K_N^{-2}\right|+\ln\left|K_N\hat{J}K_N\right|\\ &= (2n_c+1)\ln N_s+3n_c\ln N_s+\mathcal{O}(1)\\ &= (5n_c+1)\ln N_s+\mathcal{O}(1).\end{aligned}\tag{88}$$

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

$$\begin{aligned}\mathrm{BIC} &= -2\ln p_{n_c}\big(y,\hat{\theta}^{n_c}\big)+(5n_c+1)\ln N_s\\ &= 2N_s\ln\hat{\sigma}_{n_c}^2+(5n_c+1)\ln N_s\end{aligned}\tag{89}$$

where $\hat{\sigma}_{n_c}^2$ is as defined in (65), and N_s denotes the number of complex-valued data samples.

The attribute Bayesian in the name of the rule in (87) or (89) is motivated by the use of the a priori PDF $p(\theta)$, in the rule derivation, which is typical of a Bayesian approach. In fact, the BIC rule can be obtained using a full Bayesian approach, as explained next.

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

$$\begin{aligned}p(y|H_n) &= \int p_n\big(y,\theta^n\big)d\theta^n\\ &= \int p_n\big(y|\theta^n\big)p_n\big(\theta^n\big)d\theta^n.\end{aligned}\tag{90}$$

The right side of (90) is identical to that of (79). It follows from this observation and the analysis conducted in the first part of this section that, under the assumptions (77) and (78) and asymptotically in N,

$$\ln p(y|H_n)\approx\ln p_n\big(y,\hat{\theta}^n\big)-\frac{n}{2}\ln N=-\frac{1}{2}\mathrm{BIC}\tag{91}$$

[see (87)]. Hence, maximizing $p(y|H_n)$ is asymptotically equivalent with minimizing BIC, independently of the prior $p(\theta)$ [as long as it satisfies (77) and (78)]. The rediscovery of BIC in the above Bayesian framework is important, as it reveals the interesting fact that the BIC rule is asymptotically equivalent to the optimal MAP rule 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 [33] and [34], among others. In [35] and [36], the same rule has been obtained by a rather different approach based on coding arguments and the minimum description length (MDL) principle. The fact that the BIC rule can be derived in several different ways suggests that it may 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,

$$\text{For BIC}: \text{ the probability of correct detection} \rightarrow 1 \text{ as } N \rightarrow \infty \tag{92}$$

(see, e.g., [2], [3]). This should be contrasted with the nonzero overfitting probability of AIC and GIC (with $\rho =$ constant), see (67) and (68). Note that the result in (92) 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 (78) that $p(\theta)$ is independent of N, then the term $\ln p(\hat{\theta})$ can no longer be eliminated from (84) by letting $N \rightarrow \infty$. Consequently, (84) would lead to a prior-dependent rule that could be used to obtain any other rule described in this article by suitably choosing the prior. While this line of argument can serve the theoretical purpose of interpreting various rules in a Bayesian framework, it appears to have little practical value, as it can hardly be used to derive new sound order-selection rules.

Summary

We begin with the observation that all the order-selection rules discussed have a common form, i.e.,

$$-2 \ln p_n\left(y, \hat{\theta}^n\right) + \eta(n, N)n, \tag{93}$$

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

$$\begin{aligned} \text{AIC}: &\quad \eta(n, N) = 2 \\ \text{AIC}_c: &\quad \eta(n, N) = 2\frac{N}{N-n-1} \\ \text{GIC}: &\quad \eta(n, N) = \nu = \rho + 1 \\ \text{BIC}: &\quad \eta(n, N) = \ln N. \end{aligned} \tag{94}$$

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

▲Obtain an explicit expression for the term $-2 \ln p_n(y, \hat{\theta}^n)$ in (93). This requires the specification of the model structures to be tested as well as 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 article.

▲ Count the number of unknown (real-valued) parameters in each model structure under consideration. This is easily done in most parametric signal processing problems.

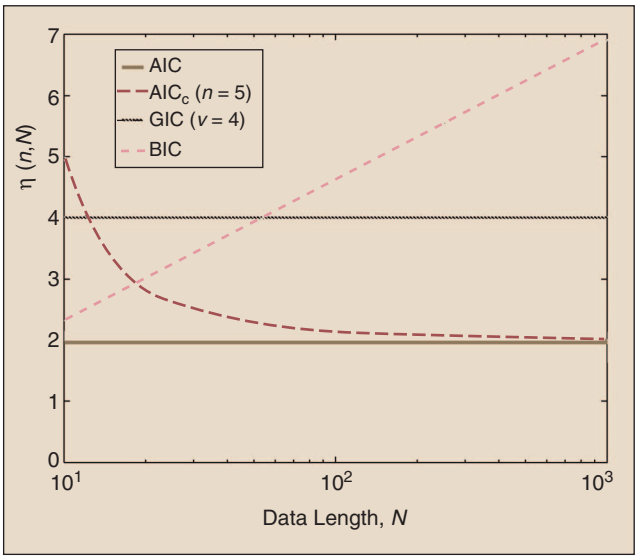
▲Verify that the assumptions that have been made to derive the rules hold true. Fortunately, the general assumptions made are quite weak and, hence, they will usually hold: indeed, the models under test may be either nested or nonnested, and they may even be only approximate descriptions of the data generating mechanism. There are two particular assumptions, made on the information matrix J, however, that do not always hold, and hence they must be checked. First, we assumed in all derivations that the inverse matrix J^{-1} exists, which 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, and a different normalization of J is required to make it tend to a constant matrix as $N \rightarrow \infty$ (this aspect is important for the BIC rule only).

We have used the sinusoidal signal model as an example to illustrate the steps above and the involved aspects.

Once the above aspects have been carefully considered, we can go on to use one of the four rules in (93) and (94) for selecting the order in our estimation problem. The question as to which rule should be used is not an easy one. In general, we prefer AIC_c over AIC: indeed, there is empirical evidence that AIC_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 AIC_c on the grounds that BIC is an asymptotic approximation of the optimal MAP rule. Regarding GIC, as mentioned earlier, GIC with $\nu \in [2, 6]$ (depending on the scenario under study) can outperform AIC and AIC_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. In summary, a possible ranking of the four rules discussed herein is as follows (the first being considered the best):

▲ BIC
▲ GIC with $\nu = 4$ ($\rho = 3$)
▲ AIC_c
▲ AIC.

We should warn the reader, however, that the previous ranking is approximate, and it will not necessarily hold in every application. In Figure 1, we show



▲ *1. Penalty coefficients of AIC, GIC with $\nu = 4$, AIC_c (for $n = 5$), and BIC as functions of the data length N.*

the penalty coefficients of the above rules, as functions of N, to further illustrate the relationship between them.

Finally, we note that in the interest of brevity we will not include numerical examples with the order-selection rules under discussion but instead refer the reader to the abundant literature on the subject; see, e.g., [1]–[6], [30]–[32]. In particular, a forthcoming article [37] contains a host of numerical examples with the information criteria discussed in this review, along with general guidelines as to how a numerical study of an order-selection rule should be organized and what performance measures should be used.

Acknowledgments

We are grateful to Prof. Randolph Moses for his assistance in the initial stages of preparing this article. This work was partly supported by the Swedish Science Council (VR). For more details on the topic of the article and its connections with other topics in signal processing, the reader may consult the book *Spectral Analysis of Signals* by P. Stoica and R. Moses (Prentice-Hall, 2004).

Petre Stoica is a professor of system modeling at Uppsala University in Sweden. For more details, see http://www.syscon.uu.se/Personnel/ps/ps.html.

Yngve Selén received the M.Sc. degree in engineering physics in 2002. He is currently a Ph.D. student of electrical engineering with specialization in signal processing at the Department of Information Technology, Uppsala University, Sweden.

References

[1] B. Choi, *ARMA Model Identification*. New York: Springer-Verlag, 1992.

[2] T. Söderström and P. Stoica, *System Identification*. London, U.K.: Prentice-Hall Int., 1989.

[3] A.D.R. McQuarrie and C.-L. Tsai, *Regression and Time Series Model Selection*. Singapore: World Scientific, 1998.

[4] H. Linhart and W. Zucchini, *Model Selection*. New York: Wiley, 1986.

[5] K.P. Burnham and D.R. Anderson, *Model Selection and Multi-Model Inference*. New York: Springer-Verlag, 2002.

[6] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics*. Tokyo, Japan: KTK Scientific, 1986.

[7] P. Stoica, P. Eykhoff, P. Janssen, and T. Söderström, "Model structure selection by cross-validation," *Int. J. Control*, vol. 43, pp. 1841–1878, 1986.

[8] T.W. Anderson, *The Statistical Analysis of Time Series*. New York: Wiley, 1971.

[9] R.J. Brockwell and R.A. Davis, *Time Series—Theory and Methods,* 2nd ed. New York: Springer-Verlag, 1991.

[10] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: Wiley, 1992.

[11] A. Papoulis, *Signal Analysis*. New York: McGraw-Hill, 1977.

[12] B. Porat, *Digital Processing of Random Signals—Theory and Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[13] M.B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Academic, 1989.

[14] L.L. Scharf, *Statistical Signal Processing—Detection, Estimation and Time Series Analysis*. Reading, MA: Addison-Wesley, 1991.

[15] C.W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[16] L. Ljung, *System Identification—Theory for the User,* 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[17] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, "Exact and large sample ML techniques for parameter estimation and detection in array processing," in *Radar Array Processing*, S. Haykin, J. Litva, and T.J. Shepherd, Eds. Berlin, Germany: Springer-Verlag, 1993, pp. 99–151.

[18] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1783–1795, Oct. 1990.

[19] H.L. Van Trees, *Optimum Array Processing* (*Detection, Estimation, and Modulation Theory,* vol. 4). New York: Wiley, 2002.

[20] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[21] P. Djuric', "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing*, vol. 46, pp. 2726–2735, 1998.

[22] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[23] S. Kullback and R.A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, Mar. 1951.

[24] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716–723, Dec. 1974.

[25] H. Akaike, "On the likelihood of a time series model," *The Statistician*, vol. 27, no. 3, pp. 217–235, 1978.

[26] R.L. Kashyap, "Inconsistency of the AIC rule for estimating the order of AR models," *IEEE Trans. Automat. Contr.*, vol. 25, no. 5, pp. 996–998, Oct. 1980.

[27] R.J. Bhansali and D.Y. Downham, "Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion," *Biometrika*, vol. 64, pp. 547–551, 1977.

[28] C. Hurvich and C. Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *J. Time Series Anal.*, vol. 14, pp. 271–279, 1993.

[29] J.E. Cavanaugh, "Unifying the derivations for the Akaike and corrected Akaike information criteria," *Statist. Probability Lett.*, vol. 33, pp. 201–208, Apr. 1977.

[30] S. de Wade and P.M.T. Broersen, "Order selection for vector autoregressive models," *IEEE Trans. Signal Processing*, vol. 51, pp. 427–433, Feb. 2003.

[31] P.M.T. Broersen, "Finite sample criteria for autoregressive order selection," *IEEE Trans. Signal Processing*, vol. 48, pp. 3550–3558, Dec. 2000.

[32] P.M.T. Broersen, "Automatic spectral analysis with time series models," *IEEE Trans. Instrum. Meas.*, vol. 51, pp. 211–216, Apr. 2002.

[33] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[34] R.L. Kashyap, "Optimal choice of AR and MA parts in autoregressive moving average models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 4, no. 2, pp. 99–104, 1982.

[35] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[36] J. Rissanen, "Estimation of structure by minimum description length," *Circuits, Syst. Signal Process.*, vol. 1, no. 4, pp. 395–406, 1982.

[37] P. Stoica and Y. Selén, "Multi-model approach to model selection," *IEEE Signal Processing Mag.*, to be published.