IT Licentiate theses
2004-003

Model Selection

Yngve Selén

UPPSALA UNIVERSITY
Department of Information Technology

UPPSALA
UNIVERSITET

Model Selection

BY
YNGVE SELÉN

October 2004

DIVISION OF SYSTEMS AND CONTROL
DEPARTMENT OF INFORMATION TECHNOLOGY
UPPSALA UNIVERSITY
UPPSALA
SWEDEN

Dissertation for the degree of Licentiate of Philosophy in Electrical Engineering:
Signal Processing
at Uppsala University 2004

Model Selection

Yngve Selén
yngve.selen@it.uu.se

Division of Systems and Control
Department of Information Technology
Uppsala University
Box 337
SE-751 05 Uppsala
Sweden

http://www.it.uu.se/

ISSN 1404-5117
Printed by the Department of Information Technology, Uppsala University, Sweden

Abstract

Before using a parametric model one has to be sure that it offers a reasonable description of the system to be modeled. If a bad model structure is employed, the obtained model will also be bad, no matter how good is the parameter estimation method. There exist many possible ways of validating candidate models. This thesis focuses on one of the most common ways, i.e., the use of information criteria. First, some common information criteria are presented, and in the later chapters, various extentions and implementations are shown. An important extention, which is advocated in the thesis, is the multi-model (or model averaging) approach to model selection. This multi-model approach consists of forming a weighted sum of several candidate models, which then can be used for inference.

Keywords: model selection; model order selection; AIC; BIC; GIC; parametric model; cross-validation; GLRT; sparse model; multi-model; model averaging; frequency-selective channel; channel averaging; sparse channel

Acknowledgments

First, I would like to express my deepest gratitude to my supervisor, Prof. Peter Stoica for wisdom, guidance, for offering me the position as a PhD student and for the occasional (i.e. an abundance of) good (well...;)) jokes. I will especially mention "Peter's Pack": Niclas Sandgren (renowned co-writer on many papers), Richard Abrahamsson and Per Åhgren for fruitful discussions and help in times of need, but also thank the rest of the Systems & Control division for general niceness and many happy times.

I would like to thank Prof. Jian Li at the University of Florida (UFL) and her students for their warm welcome during my stay in the fall of 2003. I would also like to send a special "thank you" to Dr. Erik G. Larsson from George Washington University (GWU) (but formerly of our division (once a Syscon, always a Syscon)) for lots of fun and a great welcome in Florida, and for mentoring and aiding me during the writing process of our joint papers. Another person I would like to thank is Prof. Sabine Van Huffel from Katholieke Universiteit Leuven (KUL) in Belgium for her help and guidance in my other field of research: Magnetic Resonance Spectroscopy.

Very special thanks goes to my faculty opponent, Prof. Håkan Hjalmarsson, from the Royal Institute of Technology (KTH) in Stockholm, for taking the time to read and offer constructive criticism of this thesis.

Finally, I'd like to thank all makers of wonderful music, authors of great books, brewers of fine beers, family, friends and Special ones for making life great! You all rule!

Contents

Chapter 1

Introduction

This thesis deals with the art of model selection, specifically with using Information Criteria (IC) for this purpose. The reader is assumed to possess some basic knowledge on this topic, and thus this introduction will not contain any motivation or such to this important topic in parametric modeling. For some basic information on this topic, see, e.g., [1].

1.1 Thesis outline

The chapters in this thesis are based on the papers below. They have been intentionally kept as self-contained as possible, which means that there are some (small) overlaps between them. Furthermore, the notation may sometimes vary between chapters, but is always consistent within a chapter.

Chapter 2

Here, some widely used Information Criterion (IC) rules for model selection are presented in a tutorial manner. The material in Chapter 2 is based on:

- P. Stoica and Y. Selén, "Model-Order Selection – A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, pp. 36-47, July 2004.

Chapter 3

In this chapter, two previously introduced cross-validation based rules for model order selection are presented and compared to the classical AIC and BIC/MDL rules. The material in Chapter 3 is based on:

- P. Stoica and Y. Selén, "Cross-Validation Rules for Order Estimation," *Digital Signal Processing*, vol. 14, pp. 355-371, 2004.

Chapter 4

Here, a certain equivalence between the hypothesis-theory based General Likelihood Ratio Test (GLRT) and the IC rules is shown. As a corollary, a way of exploiting the local behavior of an IC for selecting the structure of sparse models is introduced and evaluated numerically. The material in Chapter 4 is based on:

- P. Stoica, Y. Selén, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection," *IEEE Signal Processing Letters*, vol. 11, October 2004, to appear.

Chapter 5

This chapter presents the multi-model approach to model selection, in which several candidate models are estimated and combined in a weighted manner using their posterior probabilities. Via some numerical examples it is shown that this multi-model approach offers better predictive performance than the corresponding single model approaches. The material in Chapter 5 is based on:

- P. Stoica, Y. Selén, and J. Li, "Multi-model approach to model selection," *Digital Signal Processing*, 2004, in press.

Chapter 6

In this chapter the multi-model (or model averaging) approach is applied to a practical example: that of adaptive equalization for time-dispersive communication channels whose impulse responses have unknown length. Under some assumptions an optimal receiver is derived along with some computationally efficient approximations of it. Via numerical simulations these new receivers are shown to outperform conventional adaptive Viterbi equalizers that use a fixed or estimated channel length. The material in Chapter 6 is based on:

- E. G. Larsson, Y. Selén, and P. Stoica, "Adaptive equalization for frequency-selective channels of unknown length," *IEEE Transactions on Vehicular Technology*, 2004, submitted.

- E. G. Larsson, Y. Selén, and P. Stoica "Adaptive equalization for frequency-selective channels of unknown length," in *IEEE Global Telecommunications Conference (Globecom)*, (Dallas, Texas, USA), November 29 - December 3 2004, accepted for publication.

Chapter 7

In this short final chapter, a new method is presented which estimates several sparse channels and combines them in a multi-model manner, thereby combining the ideas of Chapter 4 and 5. By assuming a white training signal (for estimating the channel coefficients) this can be done in a computationally efficient way. The new method is evaluated in a communication context, similar to that in Chapter 6, and is shown to perform better, when equalizing sparse channels, than a standard single-model based equalization method and a standard multi-model based equalization method (neither of which assumes the channel to be sparse). The material in Chapter 7 is based on:

- Y. Selén, P. Stoica, and N. Sandgren, "A model averaging approach for equalizing sparse communication channels," in *Asilomar Conference on Signals, Systems, and Computers*, (Pacific Grove, California, USA), November 7-10 2004, accepted for publication.

1.2 Beside the thesis

Apart from the work presented in this thesis, I have been involved in other research projects, most notably in the field of Magnetic Resonance Spectroscopy (MRS). For conciseness reasons and for the purpose of obtaining a unified thesis, this material has been left out of the thesis, but much of it will probably be included in my forthcoming PhD thesis. The list of my papers which are not included in this thesis is as follows:

- Y. Selén, P. Stoica, N. Sandgren, and S. Van Huffel, "Using prior knowledge in SVD-based NMR spectroscopy – the ATP example," in *Proceedings of the 7th International Symposium on Signal Processing and its Applications (ISSPA)*, vol. 2, (Paris, France), pp. 33-36, July 1-4 2003.

- P. Stoica, N. Sandgren, Y. Selén, L. Vanhamme, and S. Van Huffel, "Frequency-domain method based on the singular value decomposition for frequency-selective NMR spectroscopy," *Journal of Magnetic Resonance*, vol. 165, pp. 80-88, November 2003.

- P. Stoica and Y. Selén, "Cyclic minimizers, majorization techniques, and the expectation-maximization algorithm: a refresher," *IEEE Signal Processing Magazine*, vol. 21, pp. 112-114, January 2004.

- N. Sandgren, Y. Selén, P. Stoica, and J. Li, "Parametric methods for frequency-selective MR spectroscopy – a review," *Journal of Magnetic Resonance*, vol. 168, pp. 259-272, June 2004.

- T. Laudadio, Y. Selén, L. Vanhamme, P. Stoica, P. Van Hecke, and S. Van Huffel, "Subspace-based MRS Data Quantitation of Multiplets using Prior Knowledge," *Journal of Magnetic Resonance*, vol. 168, pp. 53-65, May 2004.

- P. Stoica, Y. Selén, N. Sandgren, and S. Van Huffel, "Using prior knowledge in SVD-based parameter estimation for magnetic resonance spectroscopy–the ATP example," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1568-1578, September 2004.

- N. Sandgren, P. Stoica, and Y. Selén, "Frequency-selective SVD-based magnetic resonance spectroscopy with prior knowledge," in *Asilomar Conference on Signals, Systems, and Computers*, (Pacific Grove, California, USA), November 7-10 2004, accepted for publication.

1.3 Beyond the thesis

My future research will be more geared towards the field of Magnetic Resonance (MR); finding new MRS methods and evolving to the fields of Magnetic Resonance Imaging (MRI) and functional MRI (fMRI). The most obvious extentions to this thesis, however, would be to find new areas where the multi-model approach of Chapter 5 can be promoted, similarly to how it was introduced to the area of communications in the paper which was the basis of Chapter 6. It would also be interesting to make a more thorough evaluation of the sparse multi-model approach presented in Chapter 7.

Chapter 2

A review of the AIC, GIC and BIC rules

2.1 Introduction

The parametric (or model-based) methods of signal processing require often 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:

$$\begin{aligned} y &= \text{the vector of available data (of size } N\text{)} \\ \theta &= \text{the (real-valued) parameter vector} \\ n &= \text{the dimension of } \theta \end{aligned}$$

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:

$$\begin{aligned} y &\in \mathcal{R}^N \\ \theta &\in \mathcal{R}^n \end{aligned}$$

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, 2, 3, 4, 5, 6, 7] and the many 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 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.

2.2 Maximum Likelihood Parameter Estimation

In this section we review briefly the ML method of parameter estimation and some of its main properties that are of interest in this thesis. Let

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

The ML estimate of θ, which we denote by $\hat{\theta}$, is given by the maximizer of $p(y, \theta)$, see, e.g., [1, 8, 9, 10, 11, 12, 13, 14, 15]. Alternatively, as $\ln(\cdot)$ is a monotonically increasing function,

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

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:

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

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 (2.2) that

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

where

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

We deduce from (2.3) that

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

A simple calculation based on (2.5) shows that the ML estimates of γ and σ^2 are given by:

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

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

The corresponding value of the likelihood function is given by

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

As can be seen from (2.6), in the present case the MLM indeed reduces to the NLS.

A special case of (2.2), which we will deal with in this chapter, is the sinusoidal signal model:

$$y_c(t) = \sum_{k=1}^{n_c} \alpha_k e^{i(\omega_k t + \varphi_k)} + e(t), \quad t = 1, \ldots, N_s \tag{2.9}$$

where $\{[\alpha_k \ \omega_k \ \varphi_k]\}$ denote the amplitude, frequency and phase of the k-th 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

$$N = 2N_s \tag{2.10}$$

$$n = 3n_c + 1 \tag{2.11}$$

We will use the sinusoidal signal model in (2.9) 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., [1, 16] for a discussion about the CRB). Consequently, asymptotically in N, the pdf of $\hat{\theta}$ is given by:

$$p(\hat{\theta}) = \frac{1}{(2\pi)^{n/2}|J^{-1}|^{1/2}} e^{-\frac{1}{2}(\hat{\theta}-\theta)^T J (\hat{\theta}-\theta)} \tag{2.12}$$

where

$$J = -E\left\{\frac{\partial^2 \ln p(y, \theta)}{\partial\theta \, \partial\theta^T}\right\} \tag{2.13}$$

is the so-called (Fisher) *information matrix.*

Remark: 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, 18, 19]).

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

$$\left[-\frac{1}{N}\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\,\partial\theta^T} - \frac{1}{N}J\right] \to 0 \quad \text{as } N \to \infty \tag{2.14}$$

To motivate (2.14) for the fairly general data model in (2.2) we can argue as follows. Let us rewrite the negative log-likelihood function associated with (2.2) as (see (2.5)):

$$-\ln p(y,\theta) = \text{constant} + \frac{N}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]^2 \tag{2.15}$$

where the subindex t denotes the t-th component. From (2.15) we obtain by a simple calculation:

$$-\frac{\partial \ln p(y,\theta)}{\partial\theta} = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]\mu_t'(\gamma) \\ \frac{N}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{t=1}^{N}\left[y_t - \mu_t(\gamma)\right]^2 \end{bmatrix} \tag{2.16}$$

where

$$\mu_t'(\gamma) = \frac{\partial \mu_t(\gamma)}{\partial\gamma} \tag{2.17}$$

Differentiating (2.16) once again gives:

$$-\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\,\partial\theta^T} = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{t=1}^{N} e_t\mu_t''(\gamma) + \frac{1}{\sigma^2}\sum_{t=1}^{N}\mu_t'(\gamma)\mu_t'^T(\gamma) & \frac{1}{\sigma^4}\sum_{t=1}^{N} e_t\mu_t'(\gamma) \\ \frac{1}{\sigma^4}\sum_{t=1}^{N} e_t\mu_t'(\gamma) & -\frac{N}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{t=1}^{N} e_t^2 \end{bmatrix} \tag{2.18}$$

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

$$\mu_t''(\gamma) = \frac{\partial^2 \mu_t(\gamma)}{\partial\gamma\,\partial\gamma^T} \tag{2.19}$$

Taking the expectation of (2.18) and dividing by N, we get:

$$\frac{1}{N}J = \begin{bmatrix} \frac{1}{\sigma^2}\left(\frac{1}{N}\sum_{t=1}^{N}\mu_t'(\gamma)\mu_t'^T(\gamma)\right) & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \tag{2.20}$$

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 (2.18) 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 (2.20), which concludes the motivation of (2.14). Letting

$$\hat{J} = -\left.\frac{\partial^2 \ln p(y,\theta)}{\partial\theta\,\partial\theta^T}\right|_{\theta=\hat{\theta}} \tag{2.21}$$

we deduce from (2.14), (2.20), and the consistency of $\hat{\theta}$ that, for sufficiently large values of N,

$$\frac{1}{N}\hat{J} \approx \frac{1}{N}J = \mathcal{O}(1) \tag{2.22}$$

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 (2.20) has a finite limit, as $N \to \infty$, holds for many problems, but *not* for the sinusoidal parameter estimation problem associated with (2.9). In the latter case, (2.22) needs to be modified as follows (see, e.g., [20] and [21]):

$$K_N\hat{J}K_N \approx K_NJK_N = \mathcal{O}(1) \tag{2.23}$$

where

$$K_N = \begin{bmatrix} \frac{1}{N_s^{3/2}}I_{n_c} & 0 \\ 0 & \frac{1}{N_s^{1/2}}I_{2n_c+1} \end{bmatrix} \tag{2.24}$$

and where I_k denotes the $k \times k$ identity matrix; to write (2.24), 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.

2.3 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. 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 information criterion, which lies at the basis of another approach that can be used to derive model order selection rules.

Maximum a Posteriori (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:

$$n \in [1, \overline{n}] \tag{2.25}$$

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 AR signal with coefficients $\{a_k\}$ we can define H_n as follows:

$$H_n : \ a_n \neq 0 \text{ and } a_{n+1} = \cdots = a_{\overline{n}} = 0 \tag{2.26}$$

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 (2.11), viz.

$$n = 3n_c + 1 \tag{2.27}$$

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

$$H_{n_c} : \ \alpha_k \neq 0 \text{ for } k = 1, \ldots, n_c \text{ and } \alpha_k = 0 \text{ for } k = n_c + 1, \ldots, \bar{n}_c \tag{2.28}$$

for $n_c \in [1, \overline{n}_c]$ (with the corresponding "model order," n, being given by (2.27)).

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

$$\begin{aligned} H_1 &: \text{ the signal is a first-order AR process} \\ H_2 &: \text{ the signal is a second-order AR process} \end{aligned}$$

are nested, whereas the above H_1 and

$$H_3 : \text{ the signal consists of one sinusoid in noise}$$

are non-nested. ∎

Let

$$p_n(y|H_n) = \text{ the pdf of } y \text{ under } H_n \tag{2.29}$$

Whenever we want to emphasize the possible dependence of the pdf in (2.29) on the parameter vector of the model corresponding to H_n, we write:

$$p_n(y, \theta^n) \triangleq p_n(y|H_n) \tag{2.30}$$

Assuming that (2.29) 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:

$$p_n(H_n|y) = \frac{p_n(y|H_n)p_n(H_n)}{p(y)} \tag{2.31}$$

The maximum *a posteriori* probability (MAP) rule selects the order n (or the hypothesis H_n) that maximizes (2.31). As the denominator in (2.31) does not depend on n, the *MAP rule* is given by:

$$\max_{n\in[1,\overline{n}]} p_n(y|H_n)p_n(H_n) \tag{2.32}$$

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

$$p_n(H_n) = \frac{1}{\overline{n}}, \quad n = 1, \ldots, \overline{n} \tag{2.33}$$

in which case the MAP rule reduces to (see (2.32)):

$$\max_{n\in[1,\overline{n}]} p_n(y|H_n) \tag{2.34}$$

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

$$P_{cd} = \Pr\{[(\text{decide } H_1) \cap (H_1 = \text{true})] \cup \cdots \cup [(\text{decide } H_{\overline{n}}) \cap (H_{\overline{n}} = \text{true})]\} \tag{2.35}$$

The attribute "average" that has been attached to P_{cd} is motivated by the fact that (2.35) 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 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, $\mathcal{R}^N$, into $\overline{n}$ sets $\{S_n\}_{n=1}^{\bar{n}}$, which are such that:

$$\text{We decide } H_n \text{ if and only if } y \in S_n \tag{2.36}$$

Making use of (2.36) along with the fact that the hypotheses $\{H_n\}$ are mutually exclusive, we can write P_{cd} in (2.35) as:

$$\begin{aligned} P_{cd} &= \sum_{n=1}^{\overline{n}} \Pr\{(\text{decide } H_n) \cap (H_n = \text{true})\} \\ &= \sum_{n=1}^{\overline{n}} \Pr\{(\text{decide } H_n) | H_n\} \Pr\{H_n\} \\ &= \sum_{n=1}^{\overline{n}} \int_{S_n} p_n(y|H_n) p_n(H_n)\, dy \\ &= \int_{\mathcal{R}^N} \left[\sum_{n=1}^{\overline{n}} I_n(y) p_n(y|H_n) p_n(H_n) \right] dy \end{aligned} \tag{2.37}$$

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

$$I_n(y) = \begin{cases} 1, & \text{if } y \in S_n \\ 0, & \text{otherwise} \end{cases} \tag{2.38}$$

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 the expression (2.37) for P_{cd} imply that the MAP rule in (2.32) maximizes P_{cd}, as stated. Note that the sets $\{S_n\}$ corresponding to the MAP rule are implicitly defined via (2.32); 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 we refer to [14, 22].

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 expressed using the Kullback-Leibler (KL) information or discrepancy function (see [23]):

$$D(p_0, \hat{p}) = \int p_0(y) \ln \left[\frac{p_0(y)}{\hat{p}(y)} \right] dy \tag{2.39}$$

(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 (2.39) as:

$$D(p_0, \hat{p}) = E_0\{\ln \left[\frac{p_0(y)}{\hat{p}(y)} \right]\} = E_0\{\ln p_0(y)\} - E_0\{\ln \hat{p}(y)\} \tag{2.40}$$

Next, we prove that (2.39) possesses the properties of a suitable discrepancy function, viz.

$$\begin{aligned} &D(p_0, \hat{p}) \geq 0 \\ &D(p_0, \hat{p}) = 0 \text{ if and only if } p_0(y) = \hat{p}(y) \end{aligned} \tag{2.41}$$

To verify (2.41) we use the fact that

$$-\ln \lambda \geq 1 - \lambda \quad \text{for any } \lambda > 0 \tag{2.42}$$

and

$$-\ln \lambda = 1 - \lambda \quad \text{if and only if } \lambda = 1 \tag{2.43}$$

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

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

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 next 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 the 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, 4, 5]) lend support to the above claim.

2.4 Direct Kullback-Leibler (KL) Approach: No-Name Rule

The model-dependent part of the Kullback-Leibler (KL) discrepancy, (2.40), is given by

$$-E_0\{\ln \hat{p}(y)\} \tag{2.44}$$

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 (2.44) with respect to the model order is equivalent to *maximization* of the function:

$$I(p_0, \hat{p}) = E_0\{\ln \hat{p}(y)\} \tag{2.45}$$

which is sometimes called the relative KL information. The ideal choice for $\hat{p}(y)$ in (2.45) 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

$$\hat{p}(y) = p(y, \hat{\theta}) \tag{2.46}$$

in (2.45), which would give

$$I\left(p_0, p(y, \hat{\theta})\right) = E_0\{\ln p(y, \hat{\theta})\} \tag{2.47}$$

Because the true pdf of the data vector is unknown, we cannot evaluate the expectation in (2.47). Apparently, what we could easily do is to use the following unbiased estimate of $I(p_0, p(y, \hat{\theta}))$, instead of (2.47) itself,

$$\hat{I} = \ln p(y, \hat{\theta}) \tag{2.48}$$

However, the order selection rule that maximizes (2.48) 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 (2.48) *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 (2.45):

$$\begin{aligned}\ln p_n(y, \theta^n) \approx \ln p_n(y, \hat{\theta}^n) + (\theta^n - \hat{\theta}^n)^T \left[\left. \frac{\partial \ln p_n(y, \theta^n)}{\partial \theta^n} \right|_{\theta^n = \hat{\theta}^n} \right] \\ + \frac{1}{2}(\theta^n - \hat{\theta}^n)^T \left[\left. \frac{\partial^2 \ln p_n(y, \theta^n)}{\partial \theta^n \, \partial \theta^{nT}} \right|_{\theta^n = \hat{\theta}^n} \right] (\theta^n - \hat{\theta}^n) \triangleq \ln \hat{p}_n(y)\end{aligned} \tag{2.49}$$

Because $\hat{\theta}^n$ is the maximizer of $\ln p_n(y, \theta^n)$, the second term in (2.49) is equal to zero. Hence, we can write (see also (2.22)):

$$\ln \hat{p}_n(y) \approx \ln p_n(y, \hat{\theta}^n) - \frac{1}{2}(\theta^n - \hat{\theta}^n)^T J (\theta^n - \hat{\theta}^n) \tag{2.50}$$

According to (2.12),

$$E_0\{(\theta^n - \hat{\theta}^n)^T J (\theta^n - \hat{\theta}^n)\} = \operatorname{tr}\left[J E_0\{(\theta^n - \hat{\theta}^n)(\theta^n - \hat{\theta}^n)^T\} \right] = \operatorname{tr}[I_n] = n \tag{2.51}$$

which means that, for the choice of $\hat{p}_n(y)$ in (2.50), we have

$$I = E_0 \left\{ \ln p_n(y, \hat{\theta}^n) - \frac{n}{2} \right\} \tag{2.52}$$

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

$$\ln p_n(y, \hat{\theta}^n) - \frac{n}{2} \tag{2.53}$$

The corresponding order selection rule maximizes (2.53), or, equivalently, *minimizes*

$$\mathrm{NN}(n) = -2 \ln p_n(y, \hat{\theta}^n) + n \tag{2.54}$$

with respect to the model order n. This no-name (NN) rule can be shown to perform better than that based on (2.48), but worse than the rules presented in the next sections. Essentially, the problem with (2.54) 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 (2.54) decreases with increasing n (for nested models), whereas the second term increases. Hence, the second term in (2.54) *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 (2.54), but with a larger penalty term, and they do have better properties than (2.54). Despite this fact, we have chosen to present (2.54) 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 the present section, we motivate the multiplication with -2 in going from (2.53) to (2.54). The reason for preferring (2.54) to (2.53) is simply due to the fact that for the fairly common NLS model in (2.2) and the associated Gaussian likelihood in (2.3), $-2 \ln p_n(y, \hat{\theta}^n)$ takes on the following convenient form:

$$-2 \ln p_n(y, \hat{\theta}^n) = N \ln \hat{\sigma}_n^2 + \text{ constant} \tag{2.55}$$

(see (2.5)–(2.7)). Hence, in such a case we can replace $-2 \ln p_n(y, \hat{\theta}^n)$ in (2.54) 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 (2.54).

2.5 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 maximizing the relative KL information (see (2.45)):

$$I(p_0, \hat{p}) = E_0\{\ln \hat{p}(y)\} \tag{2.56}$$

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 (2.56), 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 (2.56) is that *the expectation in* (2.56) *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 (2.56).

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:

$$\ln \hat{p}(y) = E_x\{\ln p(y, \hat{\theta}_x)\} \tag{2.57}$$

which, when inserted in (2.56), yields:

$$I = E_y\{E_x\{\ln p(y, \hat{\theta}_x)\}\} \tag{2.58}$$

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, 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 (2.58) 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 (2.49)–(2.50), yields:

$$\begin{aligned}
\ln p(y, \hat{\theta}_x) &\approx \ln p(y, \hat{\theta}_y) + (\hat{\theta}_x - \hat{\theta}_y)^T \left[\left.\frac{\partial \ln p(y, \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_y} \right] \\
&\quad + \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)^T \left[\left.\frac{\partial^2 \ln p(y, \theta)}{\partial \theta\, \partial \theta^T}\right|_{\theta=\hat{\theta}_y} \right] (\hat{\theta}_x - \hat{\theta}_y) \\
&\approx \ln p(y, \hat{\theta}_y) - \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)^T J_y (\hat{\theta}_x - \hat{\theta}_y)
\end{aligned} \tag{2.59}$$

where J_y is the J matrix, as defined in (2.21), 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:

$$\begin{aligned}
&E_y\{E_x\{(\hat{\theta}_x - \hat{\theta}_y)^T J_y (\hat{\theta}_x - \hat{\theta}_y)\}\} \\
&\quad = E_y \left\{ E_x \left\{ \operatorname{tr} \left(J_y \left[(\hat{\theta}_x - \theta) - (\hat{\theta}_y - \theta) \right] \left[(\hat{\theta}_x - \theta) - (\hat{\theta}_y - \theta) \right]^T \right) \right\} \right\} \\
&\quad = \operatorname{tr} \left[J_y \left(J_x^{-1} + J_y^{-1} \right) \right] = 2n
\end{aligned} \tag{2.60}$$

Inserting (2.60) in (2.59) yields the following asymptotic approximation of the relative KL information in (2.58):

$$I \approx E_y\{\ln p_n(y, \hat{\theta}^n) - n\} \tag{2.61}$$

(where we have omitted the subindex y of $\hat{\theta}$ but reinstated the superindex n). Evidently, (2.61) can be estimated in an unbiased manner by

$$\ln p_n(y, \hat{\theta}^n) - n \tag{2.62}$$

Maximizing (2.62) with respect to n is equivalent to *minimizing* the following function of n:

$$\text{AIC} = -2 \ln p_n(y, \hat{\theta}^n) + 2n \tag{2.63}$$

where the acronym AIC stands for *Akaike Information Criterion* (the reasons for multiplying (2.62) by -2 to get (2.63), and for the use of the word "information" in the name given to (2.63) have been explained before, see the previous two sections).

As an example, for *the sinusoidal signal model* with n_c components (see (2.9)), AIC takes on the following form (see (2.6)–(2.11)):

$$\text{AIC} = 2N_s \ln \hat{\sigma}^2_{n_c} + 2(3n_c + 1) \tag{2.64}$$

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

$$\hat{\sigma}^2_{n_c} = \frac{1}{N_s} \sum_{t=1}^{N_s} \left| y_c(t) - \sum_{k=1}^{n_c} \hat{\alpha}_k e^{i(\hat{\omega}_k t + \hat{\varphi}_k)} \right|^2 \tag{2.65}$$

Remark: AIC can also be obtained by using the following relative KL information function, in lieu of (2.58),

$$I = E_y\{E_x\{\ln p(x, \hat{\theta}_y)\}\} \tag{2.66}$$

Note that, in (2.66), x is used for validation and y for estimation. However, the derivation of AIC from (2.66) 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, 4, 5, 6]). *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, and thus there is a true order. In such a case, analytical or 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,

$$\text{the probability of underfitting} \to 0 \tag{2.67}$$

$$\text{the probability of overfitting} \to \text{constant} > 0 \tag{2.68}$$

as $N \to \infty$ (see, e.g., [3, 26]). We can see from (2.68) 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 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 (see (2.68)) is due to the fact that the term $2n$ in (2.63) (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):

$$\mathrm{GIC} = -2\ln p_n(y, \hat{\theta}^n) + \nu n \tag{2.69}$$

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 (2.69) 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 (2.69) 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 (2.58). Interestingly, for linear regression models (given by (2.2) where $\mu(\gamma)$ is a linear function of γ), the following *corrected AIC rule*, $\mathrm{AIC_c}$, can be shown to be an *exactly* unbiased estimate of I:

$$\mathrm{AIC_c} = -2\ln p_n(y, \hat{\theta}^n) + \frac{2N}{N-n-1}n \tag{2.70}$$

(see, e.g., [28, 29]). As $N \to \infty$, $\mathrm{AIC_c} \to \mathrm{AIC}$ (as expected). However, for finite values of N the penalty term of $\mathrm{AIC_c}$ is larger than that of AIC. Consequently, in finite samples $\mathrm{AIC_c}$ has a smaller risk of overfitting than AIC, and therefore we can say that $\mathrm{AIC_c}$ 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, $\mathrm{AIC_c}$ 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, 31, 32].

2.6 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.

$$N \triangleq \mathrm{length}(y) = \rho \cdot \mathrm{length}(x), \qquad \rho \geq 1 \tag{2.71}$$

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 (2.71) 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 (2.71) we have

$$J_y = \rho J_x \tag{2.72}$$

(see, e.g., (2.20)). With this small difference, the calculations in the previous section carry over to the present case and we obtain (see (2.59)–(2.60)):

$$\begin{aligned} I &\approx E_y\{\ln p(y,\hat{\theta}_y)\} \\ &\quad - \frac{1}{2} E_y\left\{E_x\left\{\mathrm{tr}\left(J_y\left[(\hat{\theta}_x-\theta)-(\hat{\theta}_y-\theta)\right]\left[(\hat{\theta}_x-\theta)-(\hat{\theta}_y-\theta)\right]^T\right)\right\}\right\} \\ &= E_y\left\{\ln p(y,\hat{\theta}_y) - \frac{1}{2}\mathrm{tr}\left[J_y\left(\rho J_y^{-1}+J_y^{-1}\right)\right]\right\} \\ &= E_y\left\{\ln p(y,\hat{\theta}_y) - \frac{1+\rho}{2} n\right\} \end{aligned} \tag{2.73}$$

An unbiased estimate of the right side in (2.73) is given by:

$$\ln p(y, \hat{\theta}_y) - \frac{1+\rho}{2} n \tag{2.74}$$

The *generalized information criterion (GIC) rule* maximizes (2.74) or, equivalently, *minimizes*

$$\text{GIC} = -2 \ln p_n(y, \hat{\theta}^n) + (1+\rho)n \tag{2.75}$$

As expected, (2.75) reduces to AIC for $\rho = 1$. Also, note that, for a given y, the order selected by (2.75) with $\rho > 1$ is always smaller than the order selected by AIC (because the penalty term in (2.75) is larger than that in (2.63)); 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 (2.75). 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 (2.75). 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.

2.7 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 (2.56)):

$$I(p_0, \hat{p}) = E_0\{\ln \hat{p}(y)\} \tag{2.76}$$

The ideal choice of $\hat{p}(y)$ would be $\hat{p}(y) = p_n(y, \theta^n)$. However, this choice is not possible 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:

$$p(\theta) \text{ is flat around } \hat{\theta} \tag{2.77}$$

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:

$$p(\theta) \text{ is independent of } N \tag{2.78}$$

As a consequence of assumption (2.78), 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 ρ =constant), and hence with a reduced risk of overfitting.

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

$$\hat{p}(y) = E_\theta \left\{ p(y, \theta) \right\} = \int p(y, \theta) p(\theta) \, d\theta \tag{2.79}$$

Remark: In the previous sections we used a surrogate likelihood given by (see (2.57)):

$$\ln \hat{p}(y) = E_x\{\ln p(y, \hat{\theta}_x)\} \tag{2.80}$$

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

$$\hat{p}(y) = E_{\hat{\theta}_x} \left\{ p(y, \hat{\theta}_x) \right\} \tag{2.81}$$

The rule that would be obtained by using (2.81) 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 (2.81) is similar to the choice in (2.79) considered in this section, with the difference that for (2.81) the "*a priori*" pdf, $p(\hat{\theta}_x)$, depends on N. ∎

To obtain a simple asymptotic approximation of the integral in (2.79) we make use of the asymptotic approximation of $p(y, \theta)$ given by (2.49)–(2.50):

$$p(y, \theta) \approx p(y, \hat{\theta}) e^{-\frac{1}{2}(\hat{\theta}-\theta)^T \hat{J} (\hat{\theta}-\theta)} \tag{2.82}$$

which holds for θ in the vicinity of $\hat{\theta}$. Inserting (2.82) in (2.79) and using the assumption in (2.77) 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:

$$\begin{aligned}
\hat{p}(y) &\approx p(y, \hat{\theta}) p(\hat{\theta}) \int e^{-\frac{1}{2}(\hat{\theta}-\theta)^T \hat{J} (\hat{\theta}-\theta)} \, d\theta \\
&= \frac{p(y, \hat{\theta}) p(\hat{\theta}) (2\pi)^{n/2}}{|\hat{J}|^{1/2}} \underbrace{\int \frac{1}{(2\pi)^{n/2} |\hat{J}^{-1}|^{1/2}} e^{-\frac{1}{2}(\hat{\theta}-\theta)^T \hat{J} (\hat{\theta}-\theta)} \, d\theta}_{=1} \\
&= \frac{p(y, \hat{\theta}) p(\hat{\theta}) (2\pi)^{n/2}}{|\hat{J}|^{1/2}}
\end{aligned} \tag{2.83}$$

(see [21] and the references therein for the exact conditions under which the above approximation holds true). It follows from (2.76) and (2.83) that

$$\hat{I} = \ln p(y, \hat{\theta}) + \ln p(\hat{\theta}) + \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\hat{J}| \tag{2.84}$$

is an asymptotically unbiased estimate of the relative KL information. Note, however, that (2.84) 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. (2.22)):

$$\ln |\hat{J}| = \ln \left| N \cdot \tfrac{1}{N} \hat{J} \right| = n \ln N + \ln \left| \tfrac{1}{N} \hat{J} \right| = n \ln N + \mathcal{O}(1) \tag{2.85}$$

where we used the fact that $|cJ| = c^n |J|$ for a scalar c and an $n \times n$ matrix J. Using (2.85) and the fact that $p(\theta)$ is independent of N (see (2.78)) yields the following asymptotic approximation of the right side in (2.84):

$$\hat{I} \approx \ln p_n(y, \hat{\theta}^n) - \frac{n}{2} \ln N \tag{2.86}$$

The *Bayesian information criterion (BIC) rule* selects the order that maximizes (2.86), or, equivalently, *minimizes*:

$$\mathrm{BIC} = -2 \ln p_n(y, \hat{\theta}^n) + n \ln N \tag{2.87}$$

We remind the reader that (2.87) has been derived under the assumption that (2.22) 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 that (cf. (2.23)–(2.24)):

$$\begin{aligned} \ln |\hat{J}| &= \ln \left| K_N^{-2} \right| + \ln \left| K_N \hat{J} K_N \right| \\ &= (2n_c + 1) \ln N_s + 3n_c \ln N_s + \mathcal{O}(1) \\ &= (5n_c + 1) \ln N_s + \mathcal{O}(1) \end{aligned} \tag{2.88}$$

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

$$\begin{aligned} \mathrm{BIC} &= -2 \ln p_{n_c}(y, \hat{\theta}^{n_c}) + (5n_c + 1) \ln N_s \\ &= 2N_s \ln \hat{\sigma}^2_{n_c} + (5n_c + 1) \ln N_s \end{aligned} \tag{2.89}$$

where $\hat{\sigma}^2_{n_c}$ is as defined in (2.65), and N_s denotes the number of complex-valued data samples.

The attribute Bayesian in the name of the rule in (2.87) or (2.89) 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:

$$p(y|H_n) = \int p_n(y, \theta^n)\, d\theta^n = \int p_n(y|\theta^n) p_n(\theta^n)\, d\theta^n \tag{2.90}$$

The right side of (2.90) is identical to that of (2.79). It follows from this observation and the analysis conducted in the first part of this section that, under the assumptions (2.77) and (2.78) and asymptotically in N,

$$\ln p(y|H_n) \approx \ln p_n(y, \hat{\theta}^n) - \frac{n}{2} \ln N = -\frac{1}{2}\text{BIC} \tag{2.91}$$

(see (2.87)). Hence, maximizing $p(y|H_n)$ is asymptotically equivalent with minimizing BIC, independently of the prior $p(\theta)$ (as long as it satisfies (2.77) and (2.78)). 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, 34] among others. In [35, 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,

$$\text{For BIC: the probability of correct detection} \to 1 \text{ as } N \to \infty \tag{2.92}$$

(see, e.g., [1, 3]). This should be contrasted with the nonzero overfitting probability of AIC and GIC (with $\rho =$ constant), see (2.67)–(2.68). Note that the result in (2.92) 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 (2.78) that $p(\theta)$ is independent of N, then the term $\ln p(\hat{\theta})$ can no longer be eliminated from (2.84) by letting $N \to \infty$. Consequently, (2.84) would lead to a prior-dependent rule which 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.

2.8 Summary

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

$$-2\ln p_n(y, \hat{\theta}^n) + \eta(n, N)n \tag{2.93}$$

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

$$\begin{aligned}
\text{AIC}: \quad & \eta(n, N) = 2 \\
\text{AIC}_\text{c}: \quad & \eta(n, N) = 2\frac{N}{N-n-1} \\
\text{GIC}: \quad & \eta(n, N) = \nu = \rho + 1 \\
\text{BIC}: \quad & \eta(n, N) = \ln N
\end{aligned} \tag{2.94}$$

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 (2.93). 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.

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

(iii) Verify that the assumptions which 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 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 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 $N \to \infty$ (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 (2.93)–(2.94) for selecting the order in our

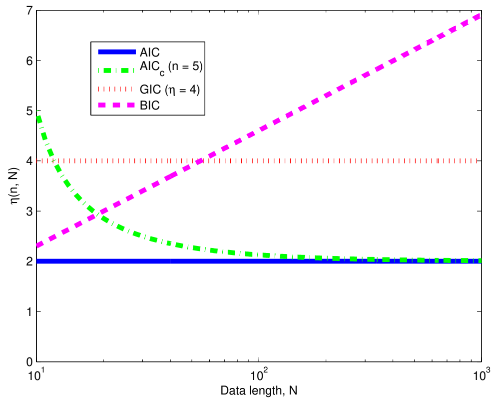


Figure 2.1: Penalty coefficients of AIC, GIC with $\nu = 4$, $\mathrm{AIC_c}$(for $n = 5$), and BIC, as functions of the data length N.

estimation problem. The question as to which rule should be used is not an easy one. In general we can prefer $\mathrm{AIC_c}$ over AIC: indeed, there is empirical evidence that $\mathrm{AIC_c}$ 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 $\mathrm{AIC_c}$ on the grounds that BIC is an asymptotic approximation of the optimal MAP rule. Regarding GIC, as mentioned earlier in this article, GIC with $\nu \in [2, 6]$ (depending on the scenario under study) can outperform AIC and $\mathrm{AIC_c}$. 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, then, a possible ranking of the four rules discussed herein is as follows (the first being considered the best):

- BIC
- GIC with $\nu = 4$ ($\rho = 3$)
- $\mathrm{AIC_c}$
- AIC

We should warn the reader, however, that the previous ranking is approximate and it will not necessarily hold in every application. In Figure 2.1 we show 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 in this chapter, but instead refer the reader to later chapters of this thesis as well as to the abundant literature on the subject, see, e.g., [1, 2, 3, 4, 5, 6, 30, 31, 32]. In particular, Chapter 5 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.

Chapter 3

Cross-validation rules for model order selection

3.1 Introduction and General Discussion

In this chapter we revisit two cross-validation rules for model order selection which appear to have some advantages over commonly-used rules, yet they do not seem to have received any significant attention so far. In an attempt to make other researchers aware of these rules, we review them here in an informal manner (with most of the technicalities in the original work being omitted) and also compare them with their closest competitors AIC, BIC (A and B Information Criteria), and MDL (Minimum Description Length) rules.

Most models dealt with in signal processing and elsewhere contain not only real-valued parameters but also integer-valued parameters. Examples of the latter parameters include the order of an AR (Auto-Regressive) equation, the length/order of a FIR (Finite Impulse Response), the number of sinusoidal components in a sinusoids-in-noise signal, the number of sources in an array output equation, the order of a Markov chain, the order of a state-space equation, and so forth. Such integer-valued parameters must be estimated along with the real-valued parameters from the available data. We will use the generic term of "order selection" for the operation of determining the integer-valued parameters of a certain model (even though sometimes "order" may not be a completely appropriate name for the integer-valued parameter in question). Note that some models may contain several integer-valued parameters. However, for the sake of notational simplicity, in what follows we consider the case of a single integer-valued parameter. Nevertheless, everything that we will say or any method that we will present can be extended mutatis mutandis to the multiple integer-valued parameter case.

Let n denote the order we want to estimate, and let $\hat{n} = 1$ be our guess for n. Evidently, this data-independent(!) guess may be rather poor, except when indeed $n = 1$ in which case the guess $\hat{n} = 1$ will be better than any other order estimate. This simple example shows that no order estimate can be uniformly (i.e., roughly speaking, in all cases) the best (i.e. the estimate that is equal to the true order most often). To obtain optimal order selection rules (i.e. rules that maximize the probability of correct estimation of n) we need to constrain the class of permissible rules.

This situation is similar to the well-known fact that a uniformly most accurate (in the mean-square error sense) method for estimating real-valued parameters does not exist in general. However, in the fairly large class of unbiased (or asymptotically unbiased) estimates the Maximum Likelihood (ML) is the optimal parameter estimation method (under regularity conditions). Furthermore, finding biased parameter estimates that outperform the ML in general is a difficult task.

In the order selection problem the situation is less fortunate. Indeed, optimal order selection rules are known only for rather particular estimation problems. Furthermore, and more important, the optimality of those rules holds only over rather limited classes of rules. More specifically, the class of permissible rules typically comprises all rules with a specified probability of overestimation (or a specified upper bound on that probability), see, e.g., [37] [38]. The problem with such a class is that in a given case there is no clear guideline for choosing the required overestimation probability (or probabilities, see [37]) to achieve a high probability of correct selection. Hence, it may easily happen that the chosen probability of overestimation is such that the probability of underestimation is rather large (even though the latter is minimized over the class of rules with the *given* probability of overestimation). The consequence is that other rules, with a different probability of overestimation (not necessarily chosen, but implicitly associated with the rule in question) may well outperform the "optimal rule", in the sense of having a higher probability of correct selection.

Owing to the difficulty of obtaining optimal rules in reasonably large classes by means of a statistical approach, researchers have considered the more pragmatic approach of developing order selection rules from *sound modeling principles* such as cross-validation (see, e.g., [39] for a general discussion). An example of a cross-validation-based rule, which we mention in passing but will not consider herein, is the PLS (Predictive Least Squares) rule [40] [41]. Many other such rules exist, based on modeling principles other than cross-validation (for an extensive review on most of the early rules, see [7]). Here we revisit the two *cross-validation-based rules* for order selection introduced in [7]. While there is no optimality claim attached to these rules, they have a number of desirable properties the most important of which is the fact that these rules make *no strong assumption on the data generating mechanism*. This fact is explained in the next section where we also clarify several aspects regarding the implementation and properties of these rules and compare them with their closest competitors (and relatives!), viz. the AIC and BIC/MDL rules.

3.2 Short-Term and Long-Term Validation Rules

Consider the following linear-regression-like model:

$$y(t) = \varphi^T(t)\theta_n + \varepsilon(t, \theta_n) \qquad t = 1, \ldots, N \tag{3.1}$$

where $\{y(t), \varphi(t)\}_{t=1}^N$ are the available data samples, θ_n is the parameter vector, n is its dimension, and $\{\varepsilon(t, \theta_n)\}$ are the residuals associated with θ_n. Both n and θ_n are unknown. For a given order n, the most commonly-used method for estimating θ_n is the Least Squares (LS):

$$\hat{\theta}_n = \arg\min_{\theta_n} V(\theta_n) \tag{3.2}$$

$$V(\theta_n) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t, \theta_n) \tag{3.3}$$

FIR and AR are two examples of well-known models that can be written in the form of (3.1) with:

$$\text{FIR:} \quad \varphi^T(t) = [u(t-1) \ldots u(t-n)] \qquad (u(t) \text{ is the input signal}) \tag{3.4}$$

$$\text{AR:} \quad \varphi^T(t) = [-y(t-1) \ldots - y(t-n)] \tag{3.5}$$

Remark: For models with memory, such as the FIR and AR in (3.4) and (3.5) above, the lower index of the sum in (3.3) (and in other similar sums appearing below) should be modified to $t = n + 1$. Hence, instead of the loss function $V(\theta_n)$ above we should use the following modified function:

$$V_n(\theta_n) = \frac{1}{N}\sum_{t=n+1}^{N} \varepsilon^2(t, \theta_n) \tag{3.6}$$

As is well known (and as can be readily checked) $\{V(\hat{\theta}_n)\}$ for $n = 1, 2, \ldots$ form a non-increasing sequence, i.e. $V(\hat{\theta}_n) \geq V(\hat{\theta}_{n+1})$. A similar property holds for (3.6):

$$V_n(\hat{\theta}_n) \geq \frac{1}{N}\sum_{t=n+2}^{N} \varepsilon^2(t, \hat{\theta}_n) \geq V_{n+1}(\hat{\theta}_{n+1}) \quad \blacksquare \tag{3.7}$$

In the following we focus on the relatively simple model class in (3.1) to simplify the discussion as much as possible. However, *the extension to a general model class is not complicated and can be found in* [7].

Let k and m be two integers such that $km = N$ (for simplicity we assume that N is exactly divisible by the chosen integers k and m, which introduces no restriction, see [7]). The choice of k (or m) will be discussed later on. Also, let (for $p = 1, \ldots, k$):

$$\begin{cases} I_p = \{(p-1)m+1, \ldots, pm\} \\ \bar{I}_p = \{1, \ldots, N\} \setminus I_p \end{cases} \tag{3.8}$$

$$\begin{cases} \bar{\theta}_n(p) & = & \arg\min_{\theta_n} \sum_{t \in \bar{I}_p} \varepsilon^2(t, \theta_n) \\ C_{ST}(n) & = & \frac{1}{N} \sum_{p=1}^{k} \sum_{t \in I_p} \varepsilon^2(t, \bar{\theta}_n(p)) \end{cases} \tag{3.9}$$

$$\begin{cases} \tilde{\theta}_n(p) & = & \arg\min_{\theta_n} \sum_{t \in I_p} \varepsilon^2(t, \theta_n) \\ C_{LT}(n) & = & \frac{1}{(k-1)N} \sum_{p=1}^{k} \sum_{t \in \bar{I}_p} \varepsilon^2(t, \tilde{\theta}_n(p)) \end{cases} \tag{3.10}$$

In the criterion $C_{ST}(n)$ above, $\bar{\theta}_n(p)$ is obtained from all subsamples except the p-th one, and the corresponding squared residuals for the remaining p-th subsample are summed-up (over p) to get the criterion. Hence, $C_{ST}(n)$ is a *short-term* (ST) *cross-validation rule* whose estimation subsample is $(k-1)$ times larger than the validation subsample. For $C_{LT}(n)$, on the other hand, the estimate $\tilde{\theta}_n(p)$ is obtained from the p-th subsample only, and its corresponding squared residuals for the remaining $(k-1)$ subsamples (that were not used for estimation) are summed-up (over p) to obtain the criterion. Consequently $C_{LT}(n)$ is a *long term* (LT) *cross-validation rule* for which the length of the estimation subsample is $(k-1)$ times smaller than that of the validation subsample.

With the above interpretation of $C_{ST}(n)$ and $C_{LT}(n)$ in mind, a sound rule for estimating n might consist of minimizing $C_{ST}(n)$ and $C_{LT}(n)$ over a given set of orders. Intuitively the order estimate obtained from $C_{ST}(n)$ should be larger than that derived from $C_{LT}(n)$ (because high-order models are better short-term than long-term extrapolators). The minimization of $C_{ST}(n)$ or $C_{LT}(n)$, however, would be computationally rather expensive. This computational difficulty issue was addressed in [7] where the following much simpler *asymptotic approximations* of $C_{ST}(n)$ and $C_{LT}(n)$ were derived:

$$\hat{C}_{ST}(n) = V(\hat{\theta}_n) + \frac{2}{N} \sum_{p=1}^{k} w_p^T \hat{R}^{-1} w_p \tag{3.11}$$

and

$$\hat{C}_{LT}(n) = V(\hat{\theta}_n) + \frac{k}{N} \sum_{p=1}^{k} w_p^T \hat{R}^{-1} w_p \tag{3.12}$$

In the above equations $\hat{\theta}_n$ is the LS estimate as defined in (3.2),

$$\hat{\theta}_n = \left[\sum_{t=1}^{N} \varphi(t)\varphi^T(t)\right]^{-1} \left[\sum_{t=1}^{N} \varphi(t)y(t)\right] \triangleq \hat{R}^{-1}\hat{r}, \tag{3.13}$$

$V(\hat{\theta}_n)$ is the corresponding LS criterion,

$$V(\hat{\theta}_n) = \frac{1}{N}\sum_{t=1}^{N} y^2(t) - \frac{\hat{r}^T\hat{R}^{-1}\hat{r}}{N} \tag{3.14}$$

and

$$w_p = \sum_{t\in I_p} \varphi(t)\varepsilon(t,\hat{\theta}_n) = \sum_{t\in I_p} \left\{\varphi(t)y(t) - \left[\varphi(t)\varphi^T(t)\right]\hat{\theta}_n\right\} \tag{3.15}$$

As already indicated in the remark following equation (3.5), for models with memory (such as (3.4) and (3.5)) the lower limit of the sums in (3.13) and (3.14) should be replaced by $t = n+1$. On the other hand, when computing the sums over I_p in (3.15) we can use the last n samples from I_{p-1} (except for I_1); proceeding in this way we can obtain these sums from the corresponding terms of $\hat{r}$ and $\hat{R}$ without *any* extra computation.

Let $\nu = [n_{min}, n_{max}]$ be the interval to which the sought order is assumed to belong. Then we can estimate the order n by one of the following two rules:

Short Term Validation (STV) Rule:

$$\hat{n}_{STV} = \arg\min_{n\in\nu} \hat{C}_{ST}(n) \tag{3.16}$$

Long Term Validation (LTV) Rule:

$$\hat{n}_{LTV} = \arg\min_{n\in\nu} \hat{C}_{LT}(n) \tag{3.17}$$

A number of comments on the STV and LTV rules follow.

C1. The criterion $\hat{C}_{LT}(n)$ is a good approximation of $C_{LT}(n)$ iff both k and m are "sufficiently large", whereas $\hat{C}_{ST}(n)$ only requires that k be "large" (m can even be equal to one), see [7]. However, intuitively it is quite clear that $\hat{C}_{ST}(n)$ with a small value of m, such as $m = 1$, may have a tendency to overestimate the "true order" (because high-order models often are good very short-term (e.g. one-step) extrapolators, as already mentioned). Consequently, in general we may also want to use $\hat{C}_{ST}(n)$, like $\hat{C}_{LT}(n)$, with an $m \gg 1$ (even though this is not required for $\hat{C}_{ST}(n)$ to be a good approximation of $C_{ST}(n)$). The choice of m (or k) is further discussed in the comment C3 below.

C2. *Neither $\hat{C}_{ST}(n)$ nor $\hat{C}_{LT}(n)$ imposes any restriction on the "true system" that generated the data $\{y(t), \varphi(t)\}$.* In particular, it is not required that the "true system" belong to the model set in (3.1). This fact appears to be a significant advantage over the many existing order selection rules that were derived under the following assumption:

$$\begin{array}{c}\text{The data was generated by equation (3.1) with } n = n^* \\ \text{(the "true order") and } \theta_{n^*} = \theta^*_{n^*} \text{ (the "true parameter} \\ \text{vector"), and moreover } \{\varepsilon(t, \theta^*_{n^*})\} \text{ is a white (usually} \\ \text{Gaussian) noise sequence.}\end{array} \tag{3.18}$$

Clearly (3.18) is a strong assumption. Interestingly, it can be shown [7] that *under (3.18)* and for $N \gg 1$ the STV rule is equivalent to the AIC rule [42]:

$$\hat{n}_{AIC} = \arg\min_{n \in \nu} \left[\ln V(\hat{\theta}_n) + \frac{2n}{N} \right] \tag{3.19}$$

whereas the LTV rule is equivalent to a form of the BIC/MDL rule ([43] [35] [33]) given by:

$$\hat{n}_{MDL} = \arg\min_{n \in \nu} \left[\ln V(\hat{\theta}_n) + \frac{nk}{N} \right] \tag{3.20}$$

(for $m = 1$ the equivalence between STV and AIC holds under more general conditions than (3.18), see [7]). Practical data, of course, does not satisfy the assumption (3.18) and with this fact in mind it is sometimes claimed that the introduction of AIC or MDL does not necessarily require (3.18) (as the derivation of most other rules do). While such a claim may be questioned, what appears to be more certain is that without (3.18) *neither* AIC *nor* MDL/BIC can be given the intuitively appealing cross-validation interpretation afforded by the aforementioned equivalence with the STV and LTV rules.

C3. To apply the STV or LTV rules we must choose k (or m). Inspired by the equivalence in (3.20) with a MDL-like rule and the fact that MDL uses $k = \ln N$ (see [43]-[44]), as well as for the lack of a better choice, we may select, for both STV and LTV:

$$k = \lfloor c \ln N \rfloor \tag{3.21}$$

where $\lfloor x \rfloor$ denotes the integer part of x and c is a constant (such as $c = 1.5$, see later). Note that for any practical value of N and c we have $k > 2$. Then if both rules are used with the same value of k (and if the model set considered has a nested structure in the sense that a model corresponding to a given order can be obtained as a special case of a model with a larger order, which is true for (3.1)) the following inequality holds (see the Appendix for a proof):

$$\hat{n}_{STV} \geq \hat{n}_{LTV} \tag{3.22}$$

Note that for the AIC and the BIC/MDL rules a similar inequality holds, viz. $\hat{n}_{AIC} \geq \hat{n}_{MDL}$, even if the model set is not nested.

C4. The cross-validation rules STV and LTV are more demanding from a computational standpoint than AIC or MDL, but only slightly. Indeed, because $\hat{R}$ and $\hat{r}$ (as well as their subsample counterparts, see the discussion following equation (3.15)) are available from the computation of $\hat{\theta}_n$, the number of additional flops required by STV or LTV, for a given value of n, is of order $O(2kn^2)$, which typically is a *small fraction* of the number of flops required to compute AIC or MDL.

C5. For n larger than the "true order" $\varepsilon(t, \hat{\theta}_n)$ will be almost independent of n and hence w_p will have a nested structure as a function of n. Because $\hat{R}$ always has such a nested structure it can readily be checked that in such a case the second term in both (3.11) and (3.12) increases with increasing n. This behavior is similar to that of AIC or MDL and it is desirable since it reduces the risk of overestimation. For n less than the "true order", on the other hand, the second term in (3.11) and (3.12) may in principle also decrease as n increases. This possible behavior of the STV and LTV criteria, which cannot occur for AIC or MDL, is *also desirable* because it may reduce the risk of underestimation. Note that for n less than the "true order", the assumption (3.18) is invalid and hence the equivalence between the STV and LTV rules, on the one hand, and AIC and MDL, on the other, does not hold. For such order values the former rules may thus *behave quite differently* from the latter.

3.3 Numerical Examples

First we consider an example in which the data is generated by a fourth-order AR equation:

$$y(t) - 1.1y(t-1) + 1.2y(t-2) - 0.9y(t-3) + 0.7y(t-4) = e(t) \tag{3.23}$$

where $e(t)$ is Gaussian white noise with zero mean and unit variance. The model class is given by (3.1) and (3.5).

An obvious way of testing the quality of a model order selection rule in this case in which the true process is included in the model set would be to consider the percentage of correctly selected order. If however the intended use of the model is taken into consideration other criteria may prevail, such as the *prediction ability* of the rule. The optimal one-step predictor for the AR process in (3.1) and (3.5) is

$$\hat{y}(t) = -a_1 y(t-1) - a_2 y(t-2) \cdots - a_n y(t-n), \tag{3.24}$$

where $[a_1 \quad a_2 \ldots a_n]^T$ is the parameter vector θ_n in (3.1). For an optimal m-step prediction, (3.24) is simply used iteratively with the true values, when unavailable, replaced with their estimates (see e.g. [45]).

We may also want to use the estimated models for spectral analysis, and with this goal in mind we can compare the mean estimated power spectral densities (PSDs) as well as the root mean square errors (RMSEs) of the PSD estimates. The PSD of an ARMA process is given by

$$\Phi(\omega) = \sigma^2 \frac{|B(e^{i\omega})|^2}{|A(e^{i\omega})|^2} \tag{3.25}$$

where $B(\cdot)$ and $A(\cdot)$ are the MA and AR polynomials, and $\omega \in [0, 2\pi]$ is the frequency variable. We use (3.25) to calculate the PSDs of the true processes (AR and ARMA, see below) as well as of the AR models.

All examples given in this section were run for several values of N but we present results only for $N = 240$, for conciseness reasons. For the STV and LTV rules, a choice of the integer parameter k was needed. Based on our empirical experience and (3.21) we selected $k = \lfloor 1.5 \ln N \rfloor$. This value of k was used in all the examples presented here ($\lfloor 1.5 \ln 240 \rfloor = 8$). For all plots presented, 1000 different realizations of the true process were used and the model order selection rules were tried on orders in the interval $\nu = [1, 40]$ ($\nu = [2, 40]$ for the sinusoidal process below). For the prediction plots, one prediction was made for each of the 1000 realizations. Then, using the true data generating process, 1000 different realizations of the true process were obtained for each prediction. The standard deviations of the predicted values from these "true" values were calculated for all 1000 predictions.

In Figure 3.1, the histograms of the orders selected by the different rules are shown. The bars at $n = 4$ represent the frequency of the correctly selected order. Observe that the LTV and MDL rules behave similarly, but the MDL rule has a slightly higher frequency of correct order selection. The STV and AIC both give very high order estimates. In Figure 3.2, the standard deviations of the prediction errors for different prediction steps are shown. As expected the standard deviations converge to a value which is equal to the standard deviation of the true process (as the predicted values converge to zero). Also note that the LTV and MDL rules offer almost identical performance and that they are superior to the STV and AIC. Figure 3.3 shows the mean PSD of the models with orders given by the different rules together with the true PSD of the process. Figure 3.4 shows the RMSE of the PSD estimates. Once again, the LTV and MDL rules perform very similarly to one another, and clearly better than the STV and AIC which both give spurious peaks and large RMSEs.

Next we consider processes which are not included in the model set. First we do this by generating data using the ARMA equation:

$$\begin{aligned} &y(t) - 1.4y(t-1) + 1.5y(t-2) - 0.8y(t-3) + 0.2y(t-4) = \\ &\quad e(t) - 1.2e(t-1) + 1.1e(t-2) - 0.9e(t-3) + 0.8e(t-4) \end{aligned} \tag{3.26}$$

We use the same performance criteria as for the AR process, and the results are shown in Figures 3.5-3.8. Note that there is no longer any "true" order for the rules to select. However, the histogram of the selected orders may still be interesting and is therefore included. The performances of the LTV and MDL are again similar to one another and better (in terms of prediction and PSD estimation quality) than the performances of the STV and AIC rules (note, however, that the STV and AIC rules provide better models of the PSD in the valley regions).

Finally, we consider data generated by the following sinusoids-in-noise process:

$$y(t) = 0.2\sin(0.3\pi t + \varphi_1) + \sin(0.51\pi t + \varphi_2) + \sin(0.53\pi t + \varphi_3) + e(t) \tag{3.27}$$

where $\{e(t)\}$ is a zero mean Gaussian white noise with variance σ^2 and the initial phases $\{\varphi_i\}$ are uniformly distributed in the interval $[0, 2\pi]$. Two values of σ^2 were considered giving the following signal-to-noise ratios ($\mathrm{SNR} = 1/\sigma^2$):

$$\begin{cases} \mathrm{SNR} &= 10\mathrm{dB} \Rightarrow \sigma^2 = 0.10 \\ \mathrm{SNR} &= 0\mathrm{dB} \Rightarrow \sigma^2 = 1.0 \end{cases}$$

For each SNR value we present the same plots as for the previous examples (see Figures 3.9-3.14), except that we omit the RMSE of the PSD that cannot be calculated for this process since the power spectrum of the sinusoids is infinite at the frequencies of the sinusoids. Also note that in the present case the AR models obtained will have some poles very close to the unit circle, and hence the prediction will converge much slower than before.

The following observations can be made by inspecting the Figures 3.9-3.14. Both the prediction and PSD estimation performance of the models provided by the STV and AIC rules are better in this case than the corresponding performance of the LTV and MDL-based models. This is at least true if the prediction horizon is not too long (the very long-term predictions given by the STV and AIC models tend to be unstable, see Figures 3.10 and 3.13), and if the few spurious peaks in the PSD spectra of the STV and AIC models (see especially Figure 3.11) are ignored. Regarding the ranking of the LTV and MDL rule, they behave quite similarly to one another for SNR = 0dB, but for SNR = 10dB the LTV rule outperforms MDL with respect to both prediction and PSD estimation performance.

3.4 Concluding Remarks

It is hard to make decisive claims about the order selection rules under discussion based on a rather limited experience. Hence we will not impose any "conclusions" on the reader, but make only some suggestions. The STV and AIC rules appear to behave in almost the same fashion. Their performance for AR and ARMA processes (the model was always AR) was rather unreliable in our experiments. On the other hand, they had a better overall performance than the LTV and MDL rules for sinusoids-in-noise processes. Regarding the latter two rules, the LTV had a slight advantage over the MDL in some cases but usually their behaviors were very similar to one another.

On a different note, the STV and LTV rules were derived virtually without making any assumption on the data generating mechanism. While this fact should give them an edge over the AIC and MDL rules (usually obtained by making rather restrictive assumptions on the process that generated the data), the latter rules do not appear to have any particular problem when used on data that cannot be exactly fitted by the assumed model.

Therefore, to conclude, which rule do we recommend? The unreliability of STV and AIC in some cases was disconcerting. Consequently we tend to suggest LTV or MDL, perhaps used with a lower value of k when high order models are expected based on the a priori information on the data at hand.

3.5 Appendix: Proof of (3.22)

Let $f(n) = V(\hat{\theta}_n)$ and $g(n) = \frac{2}{N}\sum_{p=1}^{k} w_p^T \hat{R}^{-1} w_p$. Then, by the definition of $\hat{n}_{STV}$ and $\hat{n}_{LTV}$ (see (3.16) and (3.17)) we have that:

$$\begin{cases} f(\hat{n}_{STV}) + g(\hat{n}_{STV}) \leq f(\hat{n}_{LTV}) + g(\hat{n}_{LTV}) \\ f(\hat{n}_{LTV}) + \mu g(\hat{n}_{LTV}) \leq f(\hat{n}_{STV}) + \mu g(\hat{n}_{STV}) \end{cases} \tag{3.28}$$

where $\mu = k/2 > 1$ (for $N^c \geq e^3$). It follows from (3.28) that

$$f(\hat{n}_{LTV}) - f(\hat{n}_{STV}) \leq \mu[g(\hat{n}_{STV}) - g(\hat{n}_{LTV})] \leq \mu[f(\hat{n}_{LTV}) - f(\hat{n}_{STV})] \tag{3.29}$$

and hence that

$$f(\hat{n}_{LTV}) - f(\hat{n}_{STV}) \leq \mu[f(\hat{n}_{LTV}) - f(\hat{n}_{STV})] \tag{3.30}$$

Because $\mu > 1$ the inequality in (3.30) is possible if and only if

$$f(\hat{n}_{LTV}) \geq f(\hat{n}_{STV}) \tag{3.31}$$

which in turn implies (3.22), as $f(n)$ is monotonically decreasing with increasing n (under the made assumption that the model set has a nested structure).

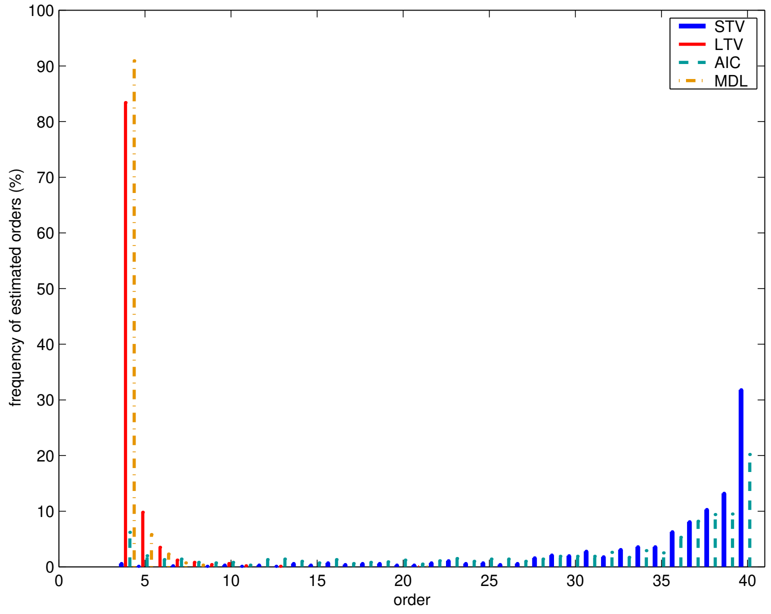


Figure 3.1: Histogram of the selected orders for the different model order selection rules. The data was given by the AR(4) process in (3.23).

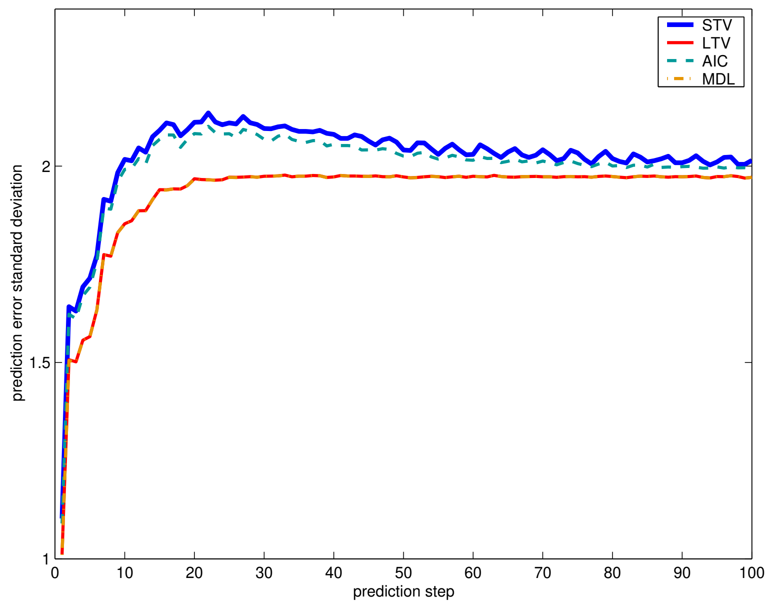


Figure 3.2: Standard deviations of the AR prediction errors for different prediction steps. The data was given by the AR(4) process in (3.23).

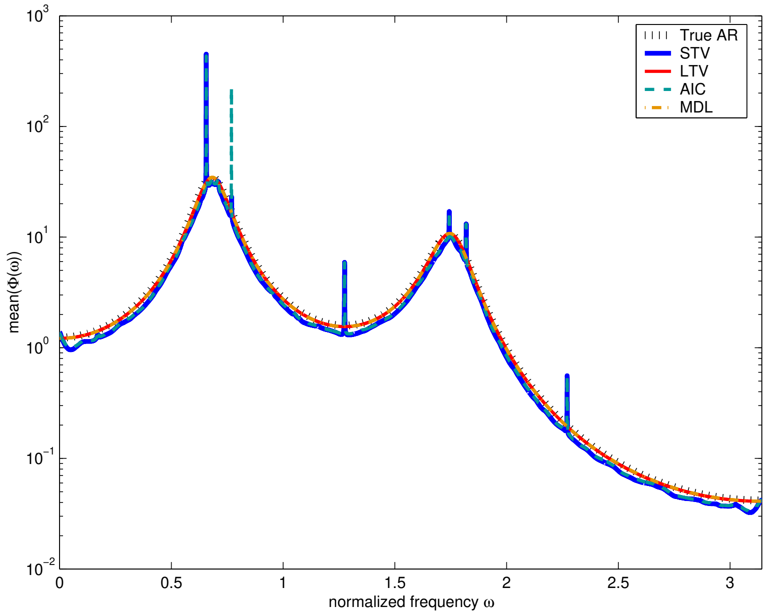


Figure 3.3: Mean PSD of the AR models. The data was given by the AR(4) process in (3.23).

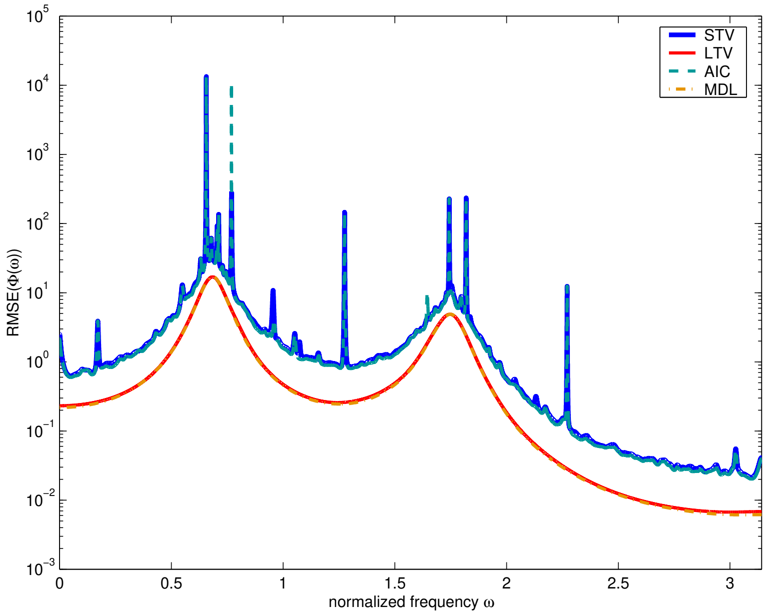


Figure 3.4: RMSE of the PSD of the AR models. The data was given by the AR(4) process in (3.23).

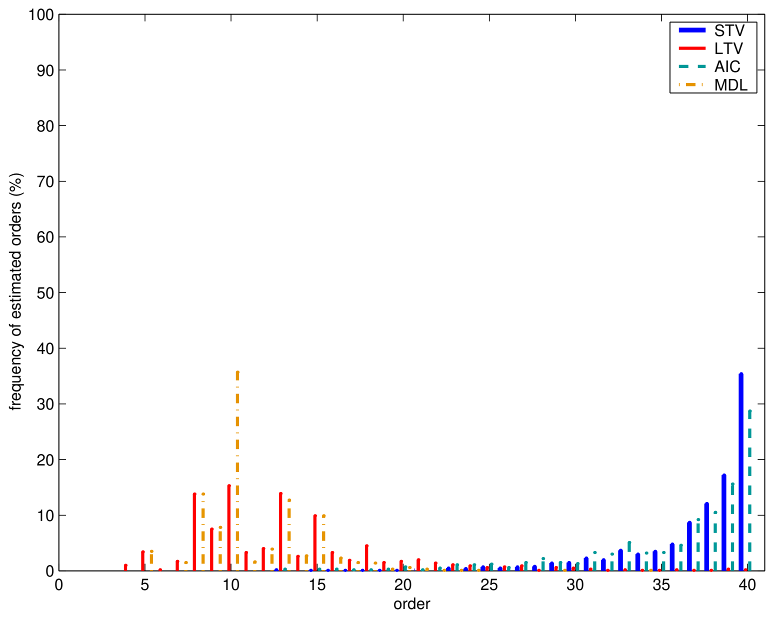


Figure 3.5: Histogram of the selected orders for the different model order selection rules. The data was given by the ARMA(4,4) process in (3.26).

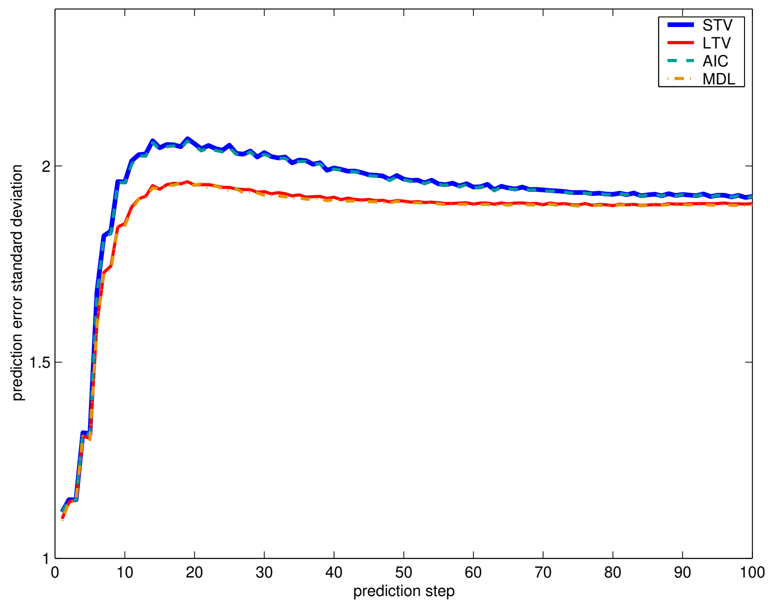


Figure 3.6: Standard deviations of the AR prediction errors for different prediction steps. The data was given by the ARMA(4,4) process in (3.26).

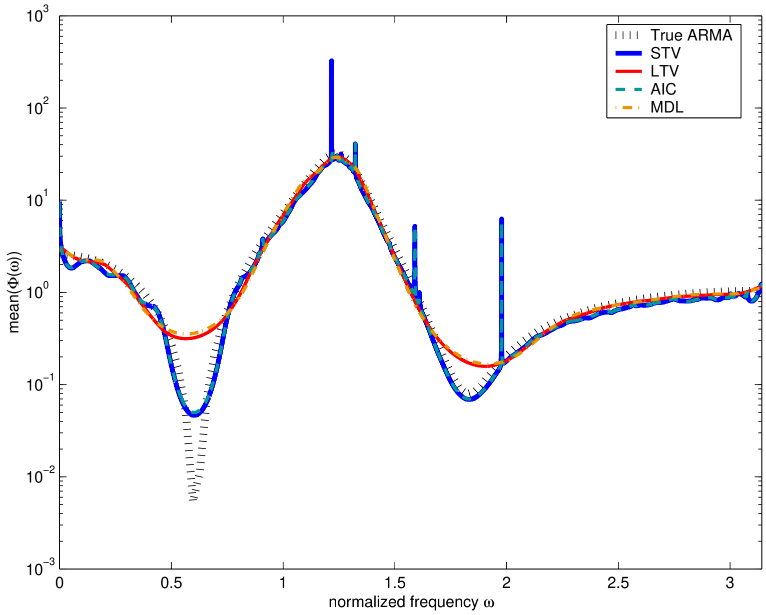


Figure 3.7: Mean PSD of the AR models. The data was given by the ARMA(4,4) process in (3.26).

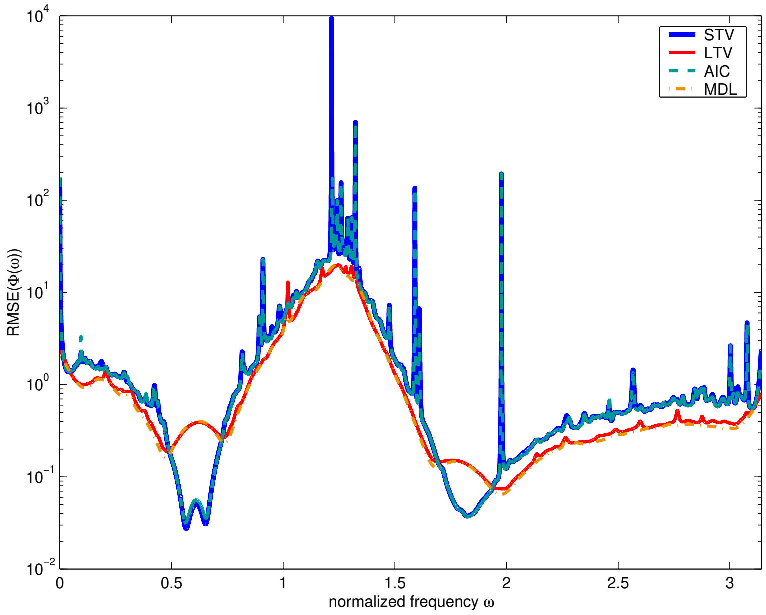


Figure 3.8: RMSE of the PSD of the AR models. The data was given by the ARMA(4,4) process in (3.26).

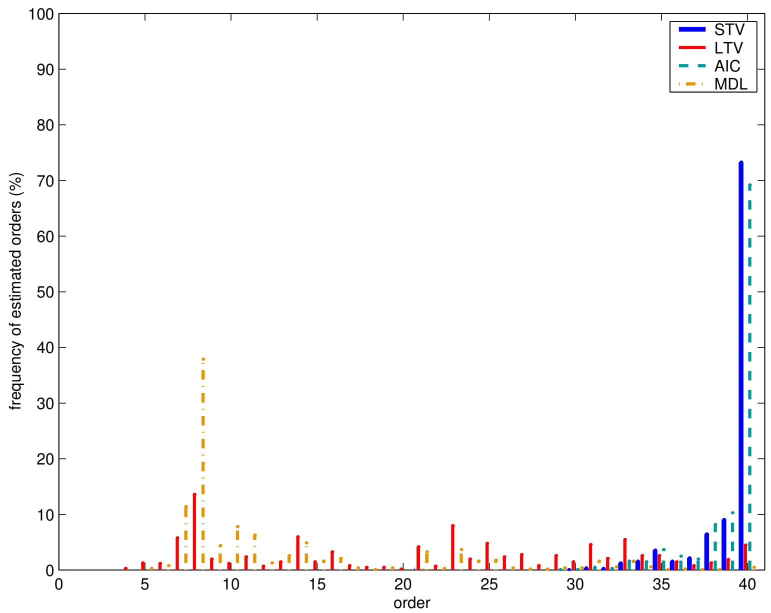


Figure 3.9: Histogram of the selected orders for the different model order selection rules. The data was given by the sinusoids-in-noise process in (3.27) with SNR = 10dB.

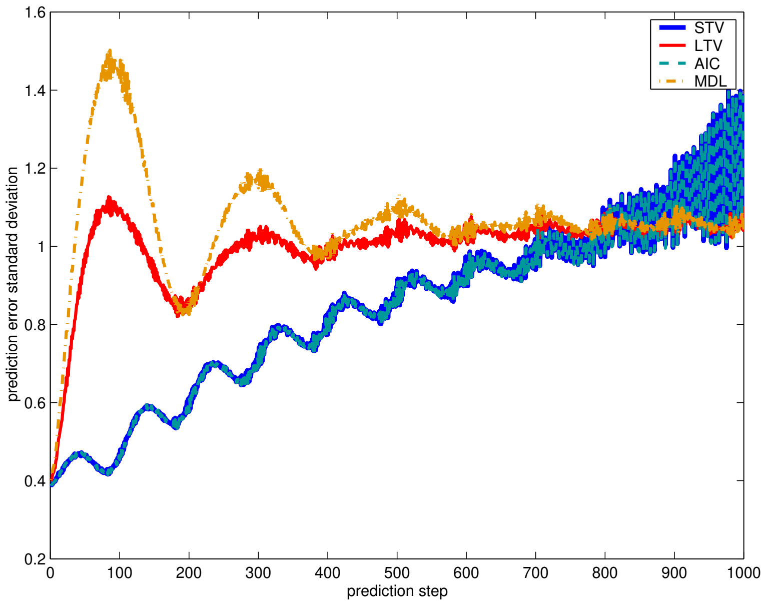


Figure 3.10: Standard deviations of the AR prediction errors for different prediction steps. The data was given by the sinusoids-in-noise process in (3.27) with SNR = 10dB.

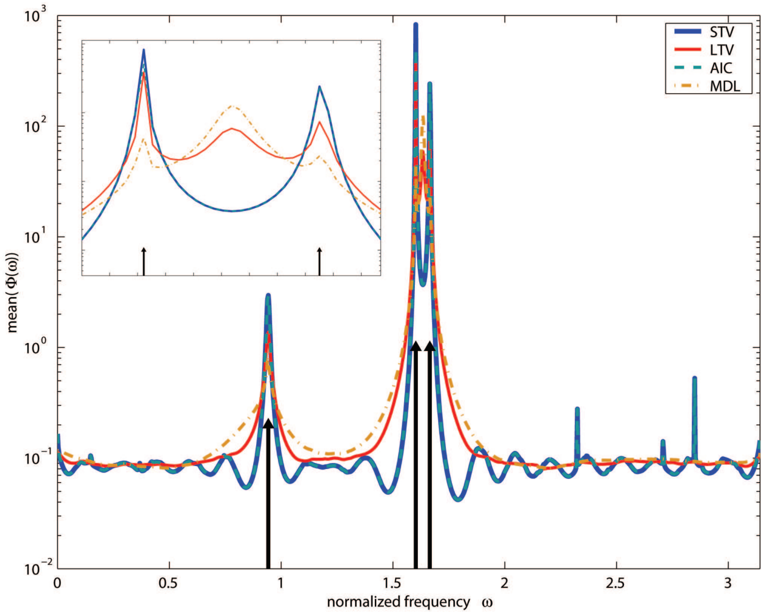


Figure 3.11: Mean PSD of the AR models together with the true line spectrum of the sinusoids. The data was given by the sinusoids-in-noise process in (3.27) with SNR = 10dB. The image in the upper left corner is a zoom of the two closely spaced peaks.

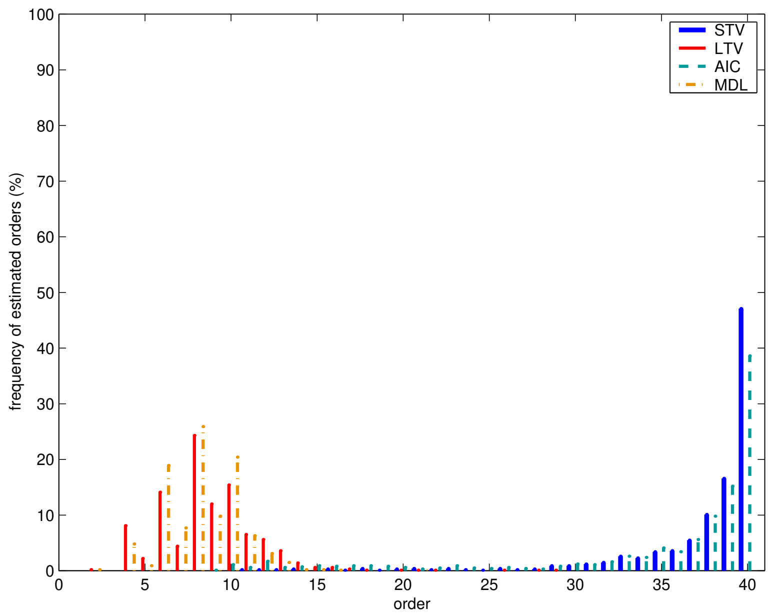


Figure 3.12: Histogram of the selected orders for the different model order selection rules. The data was given by the sinusoids-in-noise process in (3.27) with SNR = 0dB.

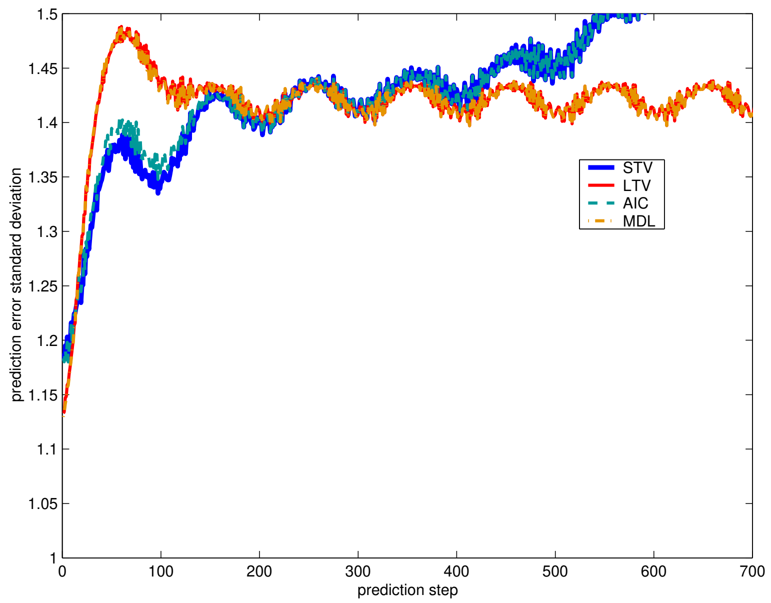


Figure 3.13: Standard deviations of the AR prediction errors for different prediction steps. The data was given by the sinusoids-in-noise process in (3.27) with SNR = 0dB.

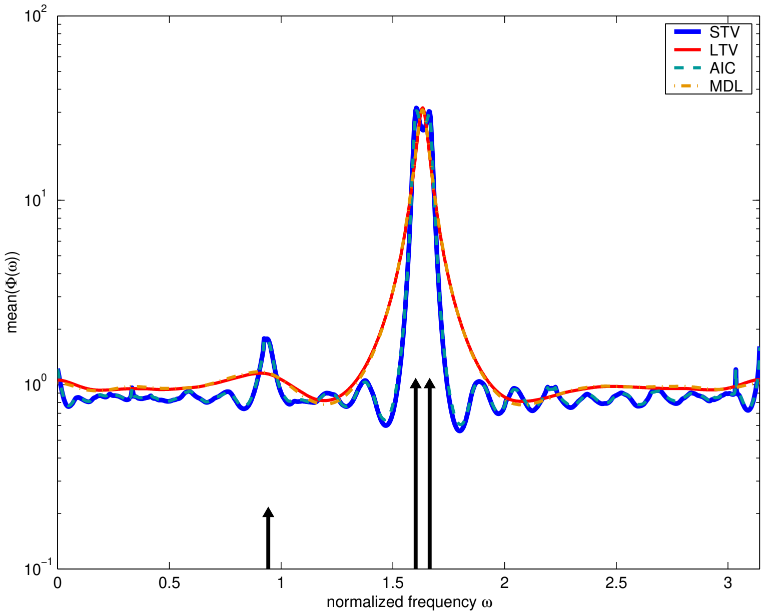


Figure 3.14: Mean PSD of the AR models together with the true line spectrum of the sinusoids. The data was given by the sinusoids-in-noise process in (3.27) with SNR = 0dB.

Chapter 4

On Information Criteria and the Generalized Likelihood Ratio Test of Model Order Selection

4.1 Introduction: The IC and the GLRT

The Information Criterion (IC) rule and the Generalized Likelihood Ratio Test (GLRT) have been usually considered to be two rather different approaches to model order selection. However, we show here that a natural implementation of the GLRT is in fact equivalent to the IC rule. A consequence of this equivalence is that a specific IC rule, such as AIC or BIC, can be viewed as a more direct way of implementing a GLRT with a specific threshold. Another consequence of the equivalence, which is emphasized herein, is a possibly original way of exploiting the information provided by the *local* behavior of an IC for selecting the structure of *sparse* models (the parameter vectors of which comprise "many" elements equal to zero).

Our discussion of the IC rule and the GLRT will necessarily be brief. Let:

$$
\begin{aligned}
\mathbf{y} \in \mathbb{R}^N \quad &= \quad \text{the available data vector} \\
\theta^n \in \mathbb{R}^n \quad &= \quad \text{the parameter vector of the } n\text{-th order model} \\
p(\mathbf{y}|\theta^n) \quad &= \quad \text{the likelihood function corresponding to the model} \\
&\qquad \text{of order } n \text{ with parameters } \theta^n \\
\hat{\theta}^n = \arg\max_{\theta^n} p(\mathbf{y}|\theta^n) \quad &= \quad \text{the maximum likelihood (ML) estimate of } \theta^n \\
n \in [n_l, n_u] \quad &= \quad \text{a given interval for the model order.}
\end{aligned}
$$

The Generalized IC (GIC) rule selects the order n as the solution to the following minimization problem:

$$\min_{n\in[n_l,n_u]}\left[\mathrm{GIC}_n \triangleq -2\ln p(\mathbf{y}|\hat{\theta}^n) + \eta \cdot n\right]. \tag{4.1}$$

The penalty factor η has different expressions for different IC rules. In particular, for the Akaike IC (AIC),

$$\eta_{\mathrm{AIC}} = 2 \tag{4.2}$$

whereas for the Bayesian IC (BIC),

$$\eta_{\mathrm{BIC}} = \ln N \quad \text{under some “regularity conditions”} \tag{4.3}$$

(see, e.g., the recent survey paper [46] and the many references therein).

The GLRT is based on the following distributional result. Let

$$\mathrm{GLRT}_{m,p} \triangleq \left[-2\ln p(\mathbf{y}|\hat{\theta}^m)\right] - \left[-2\ln p(\mathbf{y}|\hat{\theta}^p)\right] \quad \text{for } p > m. \tag{4.4}$$

Then (see, e.g., [47, 48]):

$$\mathrm{GLRT}_{m,p} \sim \chi^2(p-m) \quad (\text{for } N \gg 1, \text{ and } p > m \geq \text{true order}) \tag{4.5}$$

where $\chi^2(p-m)$ denotes the chi-square distribution with $(p-m)$ degrees of freedom. Based on (4.5), we can decide between two models with orders m and p, respectively, as follows:

$$\mathrm{GLRT}_{m,p} \underset{p}{\overset{m}{\lessgtr}} \eta(p-m) \tag{4.6}$$

where $\eta(p-m)$ is a threshold that depends on the difference $(p-m)$; we may choose $\eta(p-m)$ based on the (asymptotic) distributional result in (4.5), or possibly in a different manner (see below).

The GLRT can be used to select an order $n \in [n_l, n_u]$ in several ways. In the next section we show that a natural implementation of the GLRT for order selection is equivalent to the GIC rule. The connection between the GLRT and the GIC for comparing *two* model orders, let us say m and p, is evident from the definitions in (4.1) and (4.4), which are both based on $[-2\ln p(\mathbf{y}|\hat{\theta}^m)]$ and $[-2\ln p(\mathbf{y}|\hat{\theta}^p)]$, see [49] for details. However, here we are interested in the equivalence between the GLRT and GIC rules applied to *all* model orders in the interval $[n_l, n_u]$. Regarding this type of equivalence, it was observed in [50] that the order selected by GIC is the same as the order selected by GLRT applied to *all possible model pairs*, provided that the thresholds in the GLRT are chosen in a certain way (which was not given explicitly in [50]). The equivalence between the much more efficient implementation of GLRT presented here and GIC, however, appears to be a novel result, or at least a result that is not well known. In effect, in the previous literature authors who supported AIC or BIC usually criticized GLRT (see, e.g., [5]), and vice-versa (see, e.g., [51]).

4.2 The Main Result: The Equivalence Between GIC and GLRT

A "typical" variation of $[-2\ln p(\mathbf{y}|\hat{\theta}^n)]$, as n increases, is shown in Figure 4.1. Let

$$n_1 < n_2 < \cdots < n_s \quad \text{(for some } s \leq n_u - n_l\text{)} \tag{4.7}$$

denote the values of n for which the decrease of $[-2\ln p(\mathbf{y}|\hat{\theta}^n)]$ to the next order $(n+1)$ is "insignificant", whereas the decrease from the previous order $(n-1)$ is "significant"; see the figure. Mathematically, $n_1, \ldots, n_s$ are the values of $n \in [n_l, n_u]$ for which the following inequalities are satisfied:

$$\begin{cases} \mathrm{GLRT}_{n,n+1} < \eta & \text{(4.8a)} \\ \mathrm{GLRT}_{n-1,n} > \eta & \text{(4.8b)} \end{cases}$$

where η is a short notation for the $\eta(1)$ in (4.6). Note that for $n = n_l$ only (4.8a) is required, whereas for $n = n_u$ only (4.8b) should hold. According to (4.6), $\{n_1, \ldots, n_s\}$ are possible candidates for a GLRT-based selected order.

The next step compares the candidate models with orders $n_1, \ldots, n_s$, obtained in the previous step, with one another. Similarly to (4.8), this can be done by selecting the values of n in the set $\{n_1, \ldots, n_s\}$ for which:

$$\begin{cases} \mathrm{GLRT}_{n_k,n_{k+1}} < (n_{k+1} - n_k)\eta & \text{(4.9a)} \\ \mathrm{GLRT}_{n_{k-1},n_k} > (n_k - n_{k-1})\eta & \text{(4.9b)} \end{cases}$$

(once again, for $n = n_1$, only (4.9a) is required, whereas for $n = n_s$ only (4.9b) is necessary). Observe that the basic threshold η in (4.8) has been multiplied by $(n_{k+1} - n_k)$ and $(n_k - n_{k-1})$ in (4.9); this multiplication is suggested by the number of degrees of freedom in the distributional result (4.5).

Let $\{\bar{n}_1, \ldots, \bar{n}_{\bar{s}}\}$ be the values of n obtained in the step above. By definition, we have $\bar{s} < s$. We continue in the same fashion until we end-up with a single value of n, which is the order selected by GLRT.

Next, we note from the definitions in (4.1) and (4.4) that:

$$\mathrm{GIC}_m - \mathrm{GIC}_p = \mathrm{GLRT}_{m,p} - (p - m)\eta \tag{4.10}$$

for any positive integers p and m, with $p > m$. Assuming that the η in (4.10) (and hence in (4.1)) is the same as the η in (4.8) and (4.9), it follows from (4.10) that the inequalities in (4.9) (and therefore in (4.8), which is a special case of (4.9)) are equivalent to:

$$\begin{cases} \mathrm{GIC}_{n_k} & < & \mathrm{GIC}_{n_{k+1}} \\ \mathrm{GIC}_{n_k} & < & \mathrm{GIC}_{n_{k-1}}. \end{cases} \tag{4.11}$$

This observation allows us to deduce that: the first step of the GLRT outlined above finds all local minima of GIC_n for $n \in [n_l, n_u]$; the next step of GLRT finds all local minima of the sequence of local minimum values obtained in the first step; and so on until the last step of GLRT selects the value of n that is *the global minimum* of $\{\mathrm{GIC}_n\}$ for $n \in [n_l, n_u]$.

The equivalence between the previous implementation of GLRT and the GIC rule is thus established. As a consequence of this equivalence, the GIC rule can be viewed as a more direct way of implementing the above multi-step GLRT procedure. Furthermore, any specific GIC rule, such as AIC and BIC, can be seen as an elaborated way of choosing the threshold in the GLRT. The above equivalence also suggests a possibly original way of using GIC to select the structure of sparse models, which is described in the next section.

4.3 A Corollary: Using the GIC Rule for Sparse Models

Let θ_k^n denote the k-th element of θ^n. The parameter vector associated with a sparse model, by definition, comprises "many" elements equal to zero. For a sparse model, we may thus be interested not only in selecting the order n, but also in deciding which of the elements of the parameter vector are zero and which ones are different from zero. This means that, for such a model, we can define a *structure selection problem* (which includes the order selection) as follows: decide which elements $\{\theta_k^{n_u}\}$ of the parameter vector corresponding to the model of maximum order n_u are different from zero. This problem could in principle be solved by using the GIC rule applied to the models whose parameter vectors contain all possible combinations of zero and non-zero elements. However, even for small values of n_u the number of combinations to be tested is considerable, and the required calculations get quickly out of hand. A much simpler way of using GIC for structure selection of sparse models is suggested by the equivalence to GLRT, as explained next.

Consider, once again, the plot of $[-2\ln p(\mathbf{y}|\hat{\theta}^n)]$ versus n in Figure 4.1. This kind of variation may well occur when the data were generated by a sparse model. To see this, we first remind the reader that each decrease of -2 times the log-likelihood function which is greater than the threshold η is "significant". Consequently, if we let $\tilde{n}_G$ denote the order selected by the GLRT outlined in the previous section, and let $\tilde{n}_1, \ldots, \tilde{n}_G$ denote the orders smaller or equal to $\tilde{n}_G$ for which the decrease from the previous order is larger than η, we can decide that:

$$\theta_{\tilde{n}_1}^{n_u} \neq 0, \ldots, \theta_{\tilde{n}_G}^{n_u} \neq 0. \tag{4.12}$$

On the other hand, since any decrease of -2 times the log-likelihood function which is smaller than η is "insignificant", we can infer that:

$$\theta_k^{n_u} = 0 \quad \text{for } k \in \big\{\{n_l, \ldots, n_u\} \setminus \{\tilde{n}_1, \ldots, \tilde{n}_G\}\big\}. \tag{4.13}$$

To transfer these arguments to GIC we invoke the equivalence shown in the previous section, according to which $\tilde{n}_G$ is *the global minimum of* $\{\mathrm{GIC}_n\}$ and $\tilde{n}_1, \ldots, \tilde{n}_G$ are the *orders smaller or equal to* $\tilde{n}_G$ *for which* GIC_n *decreases from the previous order*. Then, based on (4.12) and (4.13), we can decide that:

$$\begin{cases} \theta_k^{n_u} & \neq & 0 & \text{for } k \in \{\tilde{n}_1, \ldots, \tilde{n}_G\} \\ \theta_k^{n_u} & = & 0 & \text{else.} \end{cases} \tag{4.14}$$

The solution in (4.14) to the structure selection problem, obtained by means of GIC, can be refined, if desired, by using GIC to compare it with structures obtained from it by "slight perturbations": in this way the number of structures to be tested is kept reasonably small (anyway, much smaller than the number of all possible structures). An indication of whether the structure corresponding to (4.14) is a reasonable choice can be obtained as follows. Let $\tilde{\theta}^{\tilde{n}_G}$ denote the ML estimate of $\theta^{\tilde{n}_G}$ obtained by enforcing the zero element constraints given by (4.14); and let $\widetilde{\mathrm{GIC}}_{\tilde{n}_G}$ denote the value of GIC associated with $\tilde{\theta}^{\tilde{n}_G}$. If

$$\widetilde{\mathrm{GIC}}_{\tilde{n}_G} < \mathrm{GIC}_{\tilde{n}_G} \tag{4.15}$$

then we can infer that the sparse structure in (4.14) is a good choice (note that $\widetilde{\mathrm{GIC}}_{\tilde{n}_G}$ is obtained from (4.1) using the number of *non-zero* elements in $\tilde{\theta}^{\tilde{n}_G}$ as the order n).

To illustrate the main points made in the discussion above, we present the results of a Monte-Carlo simulation study in the remainder of this section.

Numerical Example:

The data are generated by means of the equation:

$$z(t) = h_1 u(t-1) + \cdots + h_{15} u(t-15) + e(t) \quad t = 1, \ldots, N \tag{4.16}$$

where $\{z(t)\}$ denotes the output sequence, $\{u(t)\}$ is a known, white Gaussian input with mean zero and variance $\lambda^2 = 1$, $\{e(t)\}$ is a white Gaussian noise with mean zero and variance $\sigma^2 = 1$, and

$$h_3 = h_4 = h_6 = h_7 = h_8 = h_9 = h_{11} = h_{12} = h_{13} = 0. \tag{4.17}$$

The other coefficients of the sparse FIR channel in (4.16), namely $\{h_1, h_2, h_5, h_{10}, h_{14}$ and $h_{15}\}$ are independently drawn from a uniform distribution on $[-1, -0.2] \cup [0.2, 1]$. We normalize $\{h_k\}$ so that $\sum_{k=1}^{15} h_k^2 = 10$; consequently, the signal-to-noise ratio (SNR) for (4.16) is equal to

$$\mathrm{SNR} = 10 \log \left[\lambda^2 \left(\sum_{k=1}^{15} h_k^2 \right) \Big/ \sigma^2 \right] = 10 \text{ [dB]}.$$

The number of data samples assumed to be available is $N = 200$ and $N = 500$. We will make use of 10^6 realizations of the data sequence, obtained by varying the noise sequence, the input sequence, and the non-zero FIR coefficients. For each data realization we use the BIC rule, see (4.1) and (4.3), and the procedure outlined in the first part of this section, see (4.14), to select the model structure. Note that for (4.16) with $\bar{n}$ tap coefficients $\{h_1, \ldots, h_{\bar{n}}\}$ we have $\theta^n = [\sigma^2, h_1, \ldots, h_{\bar{n}}]^T$ and hence $n = \dim(\theta^n) = \bar{n} + 1$. We assume that we know that $\bar{n} \leq 20$, and hence set $n_u = 21$. Consequently, to select the model structure by using (4.14) we compute BIC_n for $n \in [1, 21]$.

Figure 4.2 shows the average results obtained from the ensemble of realizations mentioned above. As can be seen, the performance of the simple solution in (4.14) to the structure selection problem is quite satisfactory. Moreover, we note that for the structure selected with (4.14) the inequality in (4.15) was satisfied in about 87% (for $N = 200$) and 96% (for $N = 500$) of the total number of cases considered; in other words, the sparse structure selected was preferable (according to BIC) to any other full structure with $\bar{n} \in [0, 20]$ in 87% and 96% of the cases respectively.

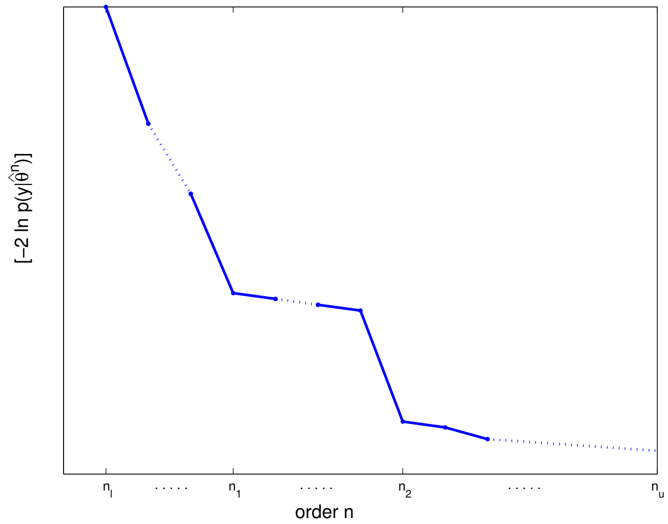
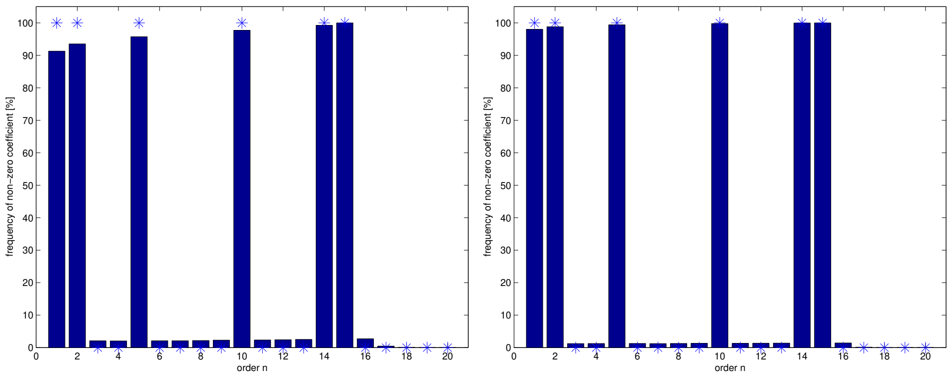


Figure 4.1: A typical variation of $[-2\ln p(\mathbf{y}|\hat{\theta}^n)]$ for increasing values of $n \in [n_l, n_u]$.



(a) $N = 200$ (b) $N = 500$

Figure 4.2: The histogram (averaged over different channel realizations) of the decision event that $h_k \neq 0$, as a function of k; the stars indicate the ideal result.

Chapter 5

Multi-model approach to model order selection

5.1 Introduction and Preliminaries

The *single-model approach* to model selection based on information criteria, such as AIC or BIC, is omnipresent in the signal processing literature. However, any single-model approach picks up only one model and hence misses the potentially significant information associated with the other models fitted to the data. In our opinion this is a drawback: indeed, depending on the application, even the true model structure (assuming that there was one) may not be the best choice for the intended use of the model. The *multi-model approach* does not suffer from such a problem: using nothing more than the values of AIC or BIC it estimates the a posteriori probabilities of each model under consideration and then it goes on to use all fitted models in a weighted manner according to their posterior likelihoods. We show via a numerical study that the multi-model approach can outperform the single-model approach in terms of statistical accuracy, without unduly increasing the computational burden. The first goal of this chapter is to advocate the multi-model approach. A second goal is to introduce some guidelines for numerically studying the performance of a model selection rule.

Almost any model used in parametric (or model-based) signal processing applications comprises both integer-valued and real-valued parameters that must be estimated from the available data. The integer-valued parameters of a model are usually referred to by the generic term of "orders". Examples of such parameters include the degree of a polynomial trend, the orders of an autoregressive moving-average equation, the dimension of a state-space equation, the number of sinusoidal components in a noisy sinusoidal signal, the number of source signals impinging on an array of sensors, and the number of coefficients in the impulse response of a linear channel.

Two well-known rules for order selection are the Akaike Information

Criterion (AIC) [24] and the Bayesian Information Criterion (BIC) [33, 34, 35]; see also [2, 1, 3, 4, 5]. These order selection rules will also be in the focus of the present chapter. Both AIC and BIC are tied to the use of the maximum likelihood (ML) method, or an asymptotic approximation thereof, for estimating the real-valued parameters of the considered models. While the ML-method is often used in signal processing applications, it is not always the method of choice and hence the dependence of AIC and BIC on the ML may seem to be a limitation. However, this is *not* necessarily so, as we explain next. Let:

$$\begin{aligned}
\mathbf{y} \in \mathbb{R}^N &= \text{the data vector} \\
\theta^n \in \mathbb{R}^n &= \text{the parameter vector of the model of order } n \\
p(\mathbf{y}|\theta^n) &= \text{the data likelihood associated with the model of order} \\
&\quad\ n \text{ with parameters } \theta^n \\
\hat{\theta}^n &= \arg\max_{\theta^n} p(\mathbf{y}|\theta^n) = \text{the ML estimate of } \theta^n \\
n \in [n_l, n_u] &= \text{a known interval for the model order } n \text{ (the indexes } l \text{ and} \\
&\quad\ u \text{ indicate the lower and upper limits, respectively)}
\end{aligned}$$

The AIC and BIC rules for model order selection are based on $\{p(\mathbf{y}|\hat{\theta}^n)\}$ for $n \in [n_l, n_u]$. However, let us assume that $\{p(\mathbf{y}|\hat{\theta}^n)\}$ are not available. This is the case, for instance, when instead of the ML estimates $\{\hat{\theta}^n\}$ we obtain parameter estimates, let us say $\{\check{\theta}^n\}$, by a different (usually simpler) method. While in such a case AIC and BIC are not directly applicable, we can derive similar order selection rules based on $\{p(\check{\theta}^{n_u}|\check{\theta}^n)\}$ instead of $\{p(\mathbf{y}|\hat{\theta}^n)\}$, assuming that the distribution function $p(\check{\theta}^{n_u}|\theta^n)$ has a known functional form (which is often the case, at least for $N \gg 1$); see [52] for this variation on the AIC and BIC theme that makes the information criterion approach to order selection *more widely applicable.*

In the signal processing literature AIC and BIC are predominantly used in a single-model selection fashion. However, the selection of a single model is often unsatisfactory since it ignores the potentially significant information associated with (some of) the other models fitted to the data. To emphasize this fact, we give examples of situations in which even selecting the true-order model is not the best choice for the intended use of the estimated model! First, consider a prediction application based on fitting a polynomial trend to $N = 20$ noisy data samples generated by a polynomial of order $n_0 = 10$. Owing to the noise present in the data, estimating 10 parameters from only 20 data samples will result in poor estimation accuracy and hence poor prediction performance; presumably, a polynomial trend based on an *underestimated* order $n < n_0$ will perform better. Next, assume data consisting of $n_0 = 2$ closely-spaced sinusoidal signals, with frequencies ω_0 and $\omega_0 + \Delta$, in strong (possibly colored) noise. In such a case many sinusoidal parameter estimation methods used with $n = 2$ will not resolve the two sinusoidal components; instead, they will provide a frequency estimate in the vicinity of ω_0 and another spurious estimate corresponding to a "noise spectral peak". However, if we apply such a method with an *overestimated* order $n > n_0$, then the two closely-spaced sinusoidal signals may be resolved and their parameters well estimated.

The multi-model approach to model selection does not suffer from the above pitfalls of the single-model approach. The multi-model approach associates posterior probabilities with each estimated model of order $n \in [n_l, n_u]$ and uses all of these models in a weighted manner according to their estimated posterior likelihoods; see, e.g., [5, 53, 54, 55, 56]. The present chapter is an attempt to advocate the multi-model approach for signal processing applications. We also introduce a methodology for running numerical performance studies of order selection rules, which appears to be sound enough for becoming a standard.

5.2 AIC and BIC Rules for Single-Model Selection

The AIC rule is usually derived using the Kullback-Leibler information ([24]), whereas the BIC rule is obtained as an asymptotic (in N) estimate of the a posteriori probability of the model given the data ([33, 34]). However, both criteria can also be obtained in a *common* cross-validation framework as briefly explained below; see Chapter 2 and the references therein for more details.

Let $\mathbf{x}$ denote a *fictitious* data sample with the same size and the same distribution function as $\mathbf{y}$, but independent of $\mathbf{y}$; and let $\tilde{\theta}^n$ denote the ML estimate of θ^n that would be obtained from $\mathbf{x}$ if $\mathbf{x}$ were available. The log-likelihood of the model of order n can then be assessed in the following cross-validatory manner:

$$E_{\tilde{\theta}^n}[\ln p(\mathbf{y}|\tilde{\theta}^n)] \tag{5.1}$$

where $\mathbf{x}$ serves as the "estimation sample" and $\mathbf{y}$ as the "validation sample", and where the expectation with respect to $\tilde{\theta}^n$ was taken to eliminate the dependence on a specific (and, anyway, unavailable) sample $\mathbf{x}$. As is known, the reason for using cross-validation is that the in-sample likelihood $p(\mathbf{y}|\hat{\theta}^n)$ is an overestimate of the "true" likelihood (owing to "noise fitting"), whereas the cross-validatory likelihood $p(\mathbf{y}|\tilde{\theta}^n)$ is a more accurate estimate. After some simple asymptotic approximations as well as taking a partial expectation with respect to $\mathbf{y}$ (see, e.g., [46]) we obtain the following estimate for (5.1):

$$\ln p(\mathbf{y}|\hat{\theta}^n) - n \tag{5.2}$$

This leads to the *AIC rule*:

$$\min_{n\in[n_l,n_u]} \left[\mathrm{AIC}_n = -2\ln p(\mathbf{y}|\hat{\theta}^n) + 2n\right] \tag{5.3}$$

which is obviously equivalent to maximizing (5.2) with respect to n.

The BIC rule can be obtained in the same cross-validation framework by making the assumption that the fictitious estimation sample $\mathbf{x}$ has now a constant length (independent of N), and using the following cross-validatory expression for the log-likelihood of the model of order n, slightly modified from (5.1)

$$\ln \left\{ E_{\tilde{\theta}^n} [p(\mathbf{y}|\tilde{\theta}^n)] \right\} \tag{5.4}$$

A derivation similar to that which leads to AIC can be used to show that the minimization of -2 times an asymptotic approximation of (5.4) yields the *BIC rule*:

$$\min_{n \in [n_l, n_u]} \left[\mathrm{BIC}_n = -2 \ln p(\mathbf{y}|\hat{\theta}^n) + (\ln N) n \right] \tag{5.5}$$

Note that we can rewrite (5.4) in a Bayesian framework:

$$\ln \left\{ \int p(\mathbf{y}|\theta^n) p(\theta^n) d\theta^n \right\} = \ln p(\mathbf{y}|M_n) \tag{5.6}$$

where M_n denotes the hypothesis that the model of order n has generated the data, and where $p(\theta^n)$ can now be interpreted as the a priori distribution function of θ^n. Since (5.6) equals (5.4) and since (5.5) is -2 times an asymptotic approximation of (5.4) (and therefore of (5.6)) it follows that *BIC_n is an asymptotic approximation (or estimate) of $-2 \ln p(\mathbf{y}|M_n)$*. From the discussion in the first part of this section it can be seen that *AIC_n can be given a similar interpretation*, with the difference that the prior distribution function of θ^n corresponding to AIC_n is dependent on N.

In the literature, the performance of AIC, BIC and other order selection rules is often discussed in terms of their probability of true-order selection. With respect to this measure of performance, BIC is superior to AIC, at least for $N \gg 1$. However, the probability of true-order selection is not the indicator of performance we are necessarily interested in: as explained in the introduction, in applications the true-order model may perform worse than other estimated models. A numerical study of the performance of AIC and BIC based on the intended use of the model will be reported in the last section.

5.3 AIC and BIC Rules for Multi-Model Selection

The following discussion concentrates on the BIC rule but its conclusions also apply to the AIC more or less directly. As explained in the previous section (see the discussion around (5.6)), for $N \gg 1$,

$$\mathrm{BIC}_n \approx -2 \ln p(\mathbf{y}|M_n) + \mathrm{const} \tag{5.7}$$

or equivalently

$$p(\mathbf{y}|M_n) \approx \text{const} \cdot e^{-\frac{1}{2}\text{BIC}_n} \tag{5.8}$$

Under the assumption that all orders $n \in [n_l, n_u]$ are a priori equiprobable (i.e. $p(M_n) = \text{const}$), we derive from (5.8) that

$$p(M_n|\mathbf{y}) = \frac{p(\mathbf{y}|M_n)p(M_n)}{p(\mathbf{y})} \approx \text{const} \cdot e^{-\frac{1}{2}\text{BIC}_n} \tag{5.9}$$

The normalizing constant in (5.9) can be obtained from the condition that

$$\sum_{k=n_l}^{n_u} p(M_k|\mathbf{y}) \approx \text{const} \cdot \sum_{k=n_l}^{n_u} e^{-\frac{1}{2}\text{BIC}_k} = 1 \tag{5.10}$$

Inserting the constant given by (5.10) into (5.9) yields:

$$\hat{p}(M_n|\mathbf{y}) = \frac{e^{-\frac{1}{2}\text{BIC}_n}}{\sum_{k=n_l}^{n_u} e^{-\frac{1}{2}\text{BIC}_k}} \tag{5.11}$$

which are *the estimated a posteriori probabilities of the hypothesized model structures* $\{M_n\}_{n=n_l}^{n_u}$ *provided by BIC.*

The BIC rule for single-model selection picks up the *mode* of $\{\hat{p}(M_n|\mathbf{y})\}$. Similarly, we can think of selecting *the median* or *the mean* of $\{\hat{p}(M_n|\mathbf{y})\}$. While comparing these possible variations of BIC with one another may be an interesting task, as explained in the introductory section we do not recommend a single-model approach for model selection. Instead, we suggest using *all* estimated models corresponding to $\{\hat{\theta}^n\}$ along with their associated estimated posterior likelihoods $\{\hat{p}(M_n|\mathbf{y})\}$, for $n = n_l, \ldots, n_u$. Presumably such a multi-model approach should be preferable to the single-model approach, especially when $\hat{p}(M_n|\mathbf{y})$ is relatively flat around its maximum with respect to n.

Results on or usages of the multi-model approach to model selection are almost inexistent in the signal processing literature; a notable exception is [57] where a multi-model-based prediction approach was proposed . On the other hand, in the statistical literature the multi-model approach is catching up with the single-model approach (see, e.g., [5, 53, 54, 55, 56]). In the latter literature the following slight *variation of the multi-model approach* is sometimes recommended. Let $n_1, \ldots, n_L$ be the values of n in the interval $[n_l, n_u]$ for which:

$$\hat{p}(M_n|\mathbf{y}) \geq \frac{\max_{n \in [n_l, n_u]} \hat{p}(M_n|\mathbf{y})}{\rho} \qquad (\text{e.g., } \rho = 10) \tag{5.12}$$

Then, instead of (5.11), we can use the following truncated (and re-normalized) posterior likelihood estimates of $\{M_n\}$:

$$\hat{p}(M_n|\mathbf{y}) = \begin{cases} \dfrac{e^{-\frac{1}{2}\mathrm{BIC}_n}}{\sum\limits_{k=1}^{L} e^{-\frac{1}{2}\mathrm{BIC}_{n_k}}} & \text{for } n = n_1, \ldots, n_L \\ 0 & \text{else} \end{cases} \tag{5.13}$$

However, if the interval $[n_l, n_u]$ was judiciously chosen then the condition in (5.12) will be satisfied for many or all of the values of n under test, and hence (5.13) will yield similar results to those obtained by (5.11). Additionally, the choice of ρ in (5.12) (5.13) is necessarily subjective[1]. With these facts in mind we will not consider the modification in (5.13) in what follows; nevertheless (5.13) may be useful whenever we expect that some model structures in the considered set $\{M_n\}_{n=n_l}^{n_u}$ are unlikely, but we do not have enough a priori information to eliminate them.

Next, let us assume that we intend to use the data model for *prediction*. Let z denote the variable we wish to predict using the available data in $\mathbf{y}$. The conditional mean of z given $\mathbf{y}$ is well known to yield the optimal (in the mean square sense) prediction:

$$\tilde{z} = E(z|\mathbf{y}) = \int z p(z|\mathbf{y}) dz \tag{5.14}$$

where the conditional probability distribution function can be written as

$$\begin{aligned} p(z|\mathbf{y}) &= \sum_{n=n_l}^{n_u} \int p(z, \theta^n, M_n|\mathbf{y}) d\theta^n \\ &= \sum_{n=n_l}^{n_u} \int p(z|\theta^n, M_n, \mathbf{y}) p(M_n|\mathbf{y}) p(\theta^n|M_n, \mathbf{y}) d\theta^n \\ &= \sum_{n=n_l}^{n_u} p(M_n|\mathbf{y}) \int p(z|\theta^n, M_n, \mathbf{y}) p(\theta^n|M_n, \mathbf{y}) d\theta^n \end{aligned} \tag{5.15}$$

Assuming that $p(\theta^n|M_n, \mathbf{y})$ is well approximated by a Dirac impulse centered at $\theta^n = \hat{\theta}^n$ (which is true for $N \gg 1$), the integral in (5.15) can be shown to be approximately equal to $p(z|\hat{\theta}^n, M_n, \mathbf{y})$. Using this approximation in (5.14) and (5.15) yields:

$$\tilde{z} \approx \sum_{n=n_l}^{n_u} p(M_n|\mathbf{y}) \int z p(z|\hat{\theta}^n, M_n, \mathbf{y}) dz = \sum_{n=n_l}^{n_u} p(M_n|\mathbf{y}) \hat{z}^n \tag{5.16}$$

where $\hat{z}^n$ is the optimal prediction obtained using the estimated model of order n (as if it were the true data generating mechanism). Replacing the unknown posterior probabilities $p(M_n|\mathbf{y})$ in (5.16) by their estimates $\hat{p}(M_n|\mathbf{y})$

[1]We note in passing that (5.13) with $\rho = 1$ gives the single-model BIC rule!

we finally obtain a *multi-model based prediction* of z in the form of a weighted average:

$$\hat{z} = \sum_{n=n_l}^{n_u} \hat{p}(M_n|\mathbf{y})\hat{z}^n \tag{5.17}$$

While $\hat{z}$ in (5.17) is only an asymptotically valid approximation of the optimal prediction $\tilde{z}$, as explained above, (5.17) has an appealing form because it requires only standard calculations to obtain $\hat{z}^n$.

For *other applications than prediction* we should proceed in a slightly different manner. In particular, we should *not* apply directly the weighted-average type of formula in (5.17) to the quantity of interest, such as a spectral density, since using the multi-model approach in that way is not guaranteed to yield better performance than a single-model approach. Instead we should make use of the multi-model approach as follows. Let the models $\{M_n\}$ be *nested* such that, for any $\bar{n} > n$, the model $M_{\bar{n}}$ with the parameter vector equal to $\begin{bmatrix} \theta^n \\ \mathbf{0} \end{bmatrix}$ reduces to the model M_n with parameter vector θ^n. In particular this means that the $n_u \times 1$ vectors $\begin{bmatrix} \hat{\theta}^n \\ \mathbf{0} \end{bmatrix}$ for $n \in [n_l, n_u]$ can all be viewed as estimates of θ^{n_u}. Similarly to the equation (5.14) for prediction, the optimal (in the mean square sense) estimate of θ^{n_u} is given by the mean of the a posteriori distribution $p(\theta^{n_u}|\mathbf{y})$:

$$\begin{aligned}
\tilde{\theta}^{n_u} &= \int \theta^{n_u} p(\theta^{n_u}|\mathbf{y}) d\theta^{n_u} \\
&= \int \theta^{n_u} \left[\sum_{n=n_l}^{n_u} p(\theta^{n_u}, M_n|\mathbf{y}) \right] d\theta^{n_u} \\
&= \sum_{n=n_l}^{n_u} p(M_n|\mathbf{y}) \int \theta^{n_u} p(\theta^{n_u}|M_n, \mathbf{y}) d\theta^{n_u} \\
&\approx \sum_{n=n_l}^{n_u} p(M_n|\mathbf{y}) \begin{bmatrix} \hat{\theta}^n \\ \mathbf{0} \end{bmatrix}
\end{aligned} \tag{5.18}$$

where, like in (5.16), the approximate equality is motivated by the fact that, for $N \gg 1$, the function $p(\theta^{n_u}|M_n, \mathbf{y})$ is well approximated by a Dirac impulse centered at $\theta^{n_u} = \begin{bmatrix} \hat{\theta}^n \\ \mathbf{0} \end{bmatrix}$. Replacing $\{p(M_n|\mathbf{y})\}$ in (5.18) by their estimates $\{\hat{p}(M_n|\mathbf{y})\}$ provided by BIC (or AIC) yields the following asymptotically optimal estimate of θ^{n_u}:

$$\hat{\hat{\theta}}^{n_u} = \hat{p}(M_{n_l}|\mathbf{y}) \begin{bmatrix} \hat{\theta}^{n_l} \\ \mathbf{0} \end{bmatrix} + \hat{p}(M_{n_l+1}|\mathbf{y}) \begin{bmatrix} \hat{\theta}^{n_l+1} \\ \mathbf{0} \end{bmatrix} + \cdots + \hat{p}(M_{n_u}|\mathbf{y})\hat{\theta}^{n_u} \tag{5.19}$$

The use of M_{n_u} with the multi-model based parameter estimate $\hat{\hat{\theta}}^{n_u}$ should outperform (for $N \gg 1$) the use of M_{n_u} with any single-model based parameter estimate $\begin{bmatrix} \hat{\theta}^n \\ \mathbf{0} \end{bmatrix}$ $(n \leq n_u)$, or equivalently, the use of M_n with any $\hat{\theta}^n$, for any $n \in [n_l, n_u]$.

Remark: Note that the use of M_{n_u} with $\hat{\hat{\theta}}^{n_u}$ for prediction leads to (5.17) whenever the predictors are linear in the model parameters. However, when the predictor is a nonlinear function of the model parameters, the use of (5.17) and (5.19) may yield different prediction results in finite samples, but the differences between these results should be small for $N \gg 1$.

5.4 Numerical Study of Performance

5.4.1 Performance Measures

Most studies of model selection rules reported in the literature make use of data generated using only one model $M_{n_0} \in \{M_n\}$, in a multiple realization manner. A possible concern with this type of approach is the fact that the data generating mechanism is assumed to belong to the class of models considered. However this is not the concern we want to discuss here. In our opinion a major concern with the aforementioned numerical approach to performance study is the use of only one (or a few) data generating mechanism(s). This is not satisfactory in several respects. First, if we choose n_0 closer to n_l than to n_u then we favor the selection rules that underestimate the order; and vice-versa if n_0 is chosen closer to n_u than to n_l. Hence we should generate data sets using all order values n in $[n_l, n_u]$. Second, using only one data generating mechanism, for each order n, may lead to biased conclusions. Consequently, for each order n in $[n_l, n_u]$ we should use as many data generating mechanisms as possible.

Note that the idea of considering all values of n in $[n_l, n_u]$ in a performance study is also supported by the classical hypothesis testing theory where the performance and optimality of a rule are defined with respect to the *total* (or *average*) probability of correct selection, which using our notation is given by (see, e.g., [46, 14, 22]):

$$p(\text{decide } M_{n_l}|M_{n_l})p(M_{n_l}) + \cdots + p(\text{decide } M_{n_u}|M_{n_u})p(M_{n_u}) \tag{5.20}$$

If the model structures considered are a priori equiprobable, as assumed before, then (5.20) reduces to:

$$\frac{1}{n_u - n_l + 1} \sum_{n=n_l}^{n_u} p(\text{decide } M_n|M_n) \tag{5.21}$$

The point to note is that it is the *total probability* (5.20), or the *average probability* (5.21), which is of interest, and not $p(\text{decide } M_n|M_n)$ for only one value of the order in the assumed interval $[n_l, n_u]$.

In summary, we should generate the data as follows: given N, n_u, and n_l, for each n in $[n_l, n_u]$ draw n_m model parameter vectors $\{\theta^n\}$ from a random distribution, and for each θ^n obtain via replication n_r random data realizations of length N (by varying the "noise" part of the data generating mechanism). In this way we obtain

$$(n_u - n_l + 1) \times n_m \times n_r \text{ data sets} \tag{5.22}$$

and we should determine the average performance of the order selection rule under analysis over the ensemble of these data sets.

The "average performance" mentioned in the previous paragraph should be defined with respect to the objective of the model. For instance, if the model is to be used for prediction then we can utilize the *average prediction mean square error* (MSE):

$$\frac{1}{(n_u - n_l + 1)n_m n_r} \sum_n \sum_m \sum_r [\hat{z} - z]^2_{n,m,r,k} \qquad \text{for } k = 1, 2, \ldots \tag{5.23}$$

where n_m and n_r have been defined before, and where $(\hat{z} - z)_{n,m,r,k}$ denotes the k-step prediction error corresponding to the r-th realization ($r = 1, \ldots, n_r$) of the data set generated by the m-th mechanism ($m = 1, \ldots, n_m$) of order n ($n = n_l, \ldots, n_u$). Whenever the objective of the model is not yet determined (or perhaps not well defined) or when the model order is a quantity of interest per se we can use the *average frequency of various selected orders* as a possible performance measure (see, however, the introductory section for a cautionary note about using this measure):

$$\frac{1}{(n_u - n_l + 1)n_m n_r} \sum_n \sum_m \sum_r \delta_{n+k, \hat{n}_{n,m,r}} \qquad \text{for } k = 0, \pm 1, \pm 2, \ldots \tag{5.24}$$

where $\hat{n}_{n,m,r}$ denotes the order estimated from the data set corresponding to (n, m, r), and $\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

5.4.2 Numerical Examples

We consider a linear channel described by the equation:

$$y(k) = h_0 u(k) + \cdots + h_n u(k - n) + e(k) \tag{5.25}$$

where $\{e(k)\}$ is a Gaussian white noise and $\{u(k)\}$ is a known white Gaussian training sequence. The coefficients $\{h_k\}$ are unknown and so is the order n of the channel. For the finite-impulse-response model in (5.25), a simple calculation shows that (see, e.g., [2, 1, 3, 4, 5]):

$$-2 \ln p(\mathbf{y}|\hat{\theta}^n) = N \ln \hat{\sigma}_n^2 + \text{const} \tag{5.26}$$

where

$$\hat{\sigma}_n^2 = \frac{1}{N} \sum_{k=1}^{N} [y(k) - \varphi^T(k)\hat{\theta}^n]^2$$

and where

$$\begin{aligned}\varphi^T(k) &= [u(k)\cdots u(k-n)]\\ \hat{\theta}^n &= \left[\sum_{k=1}^{N}\varphi(k)\varphi^T(k)\right]^{-1}\left[\sum_{k=1}^{N}\varphi(k)y(k)\right]\end{aligned}$$

Furthermore for (5.25) the prediction of $\{y(k)\}_{k=N+1,N+2,\ldots}$ based on the estimated model of order n (assumed to be the true data generating mechanism) is given by:

$$\hat{y}(k) \;=\; \hat{h}_0 u(k)+\cdots+\hat{h}_n u(k-n), \qquad k=N+1,N+2,\ldots$$

We consider the following values of N, n_l, n_u, n_m and n_r:

$$\begin{array}{rclcrcl} N &=& 100 & \text{and} & N &=& 400\\ n_l &=& 1 & & n_u &=& 10\\ n_m &=& 100 & & n_r &=& 100\end{array} \tag{5.28}$$

For each data model, the $n_r = 100$ noise and training sequences are drawn from two independent Gaussian distributions with means equal to zero and variances equal to 1 (for the training sequence) and to the values corresponding to the SNRs below (for the noise); here SNR denotes the signal-to-noise ratio, for which we consider the following two values:

$$\text{SNR} = 0\text{dB and SNR} = 10\text{dB} \tag{5.29}$$

Each tap coefficient in the $n_m = 100$ sets of coefficients $\{h_k\}$ is independently drawn from a uniform distribution in the interval $[-1, 1]$. To maintain a constant SNR, see (5.29), we normalize each vector $\mathbf{h} = \{h_k\}$ generated as above so that $\|\mathbf{h}\| = 1$.

Figure 5.1 shows the histograms of the orders selected by AIC and BIC for the four combinations of N and SNR in (5.28) and (5.29); see (5.24) for the definition of the histogram. In this figure we also show the average posterior probability estimates obtained using AIC and BIC; see equation (5.11) for BIC; the corresponding formula for AIC is similar. We can see from Figure 5.1 that, as expected, AIC overestimates the true order more often than BIC, and that BIC underestimates the true order more often than AIC; also BIC selects the true order most often. Note that, among other things, the histograms of estimated orders are useful in evaluating the Cramér-Rao bounds and the estimation accuracy of the parameter estimates when not only the real-valued parameters but also the orders are estimated, see e.g. [58].

Figure 5.2 shows the corresponding average prediction MSE curves (see (5.23) for a definition). The differences between the performances of the various model selection rules are small, which appears to be due to the fact

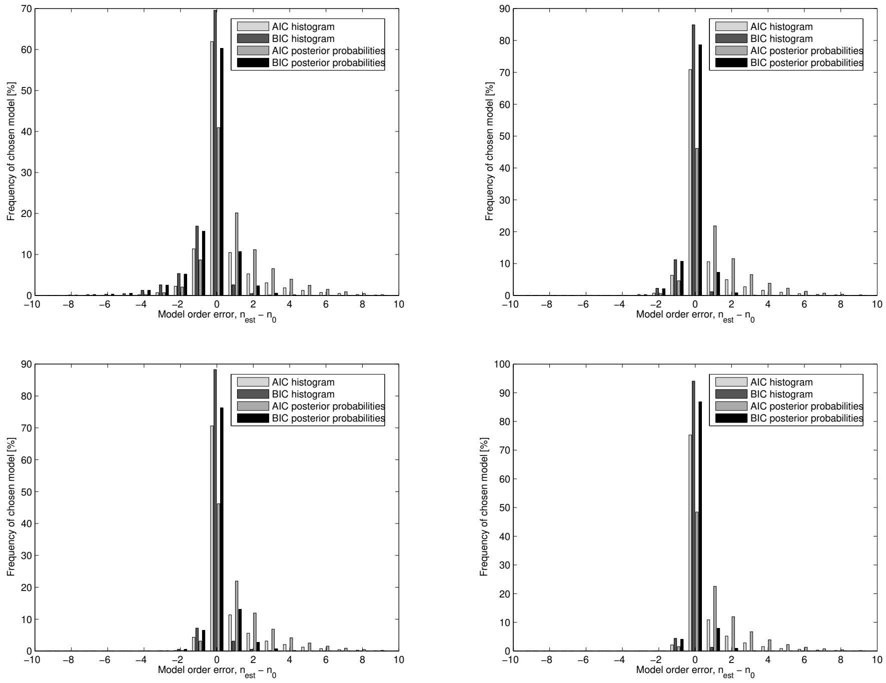


Figure 5.1: Histograms and averaged posterior probabilities of estimated orders for $N = 100$ (left), $N = 400$ (right), SNR = 0dB (upper) and SNR = 10dB (lower).

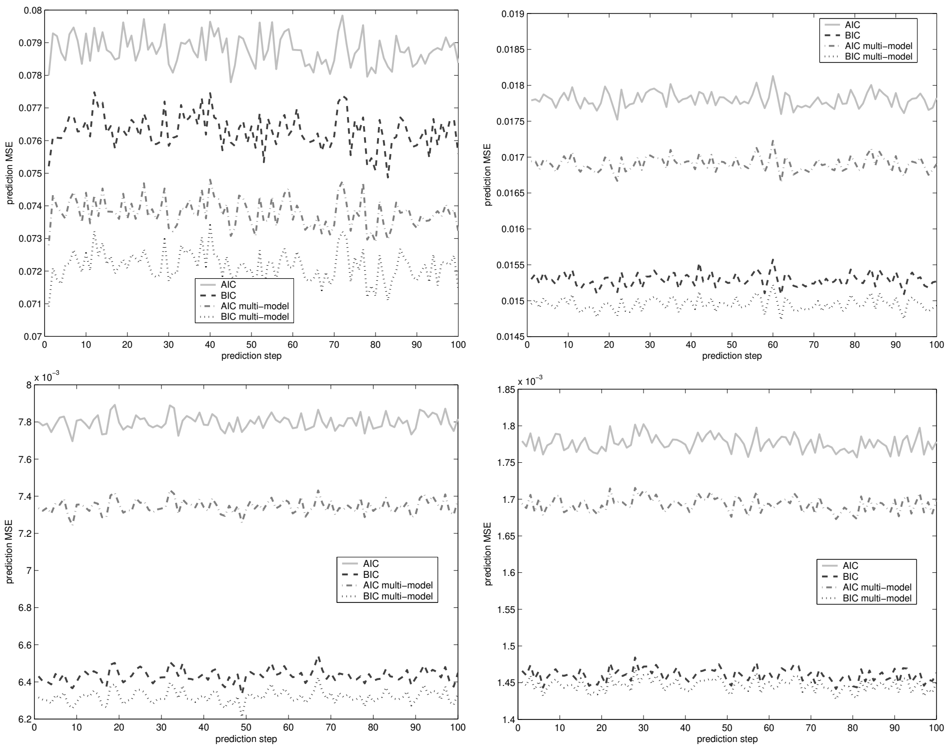


Figure 5.2: Average prediction MSEs for $N = 100$ (left), $N = 400$ (right), SNR $= 0$dB (upper) and SNR $= 10$dB (lower).

that in the present case the true order has a much higher posterior likelihood than the other orders, especially for BIC (see Figure 5.1). However, despite being relatively small, the differences between the various model selection rules are consistent (the curves in Figure 5.2 do not intersect). Note from the figure that, as expected, the multi-model approach outperforms the single-model approach.

Chapter 6

Adaptive equalization for frequency-selective channels of unknown length

6.1 Introduction

This chapter studies adaptive equalization for time-dispersive communication channels whose impulse responses have unknown lengths. This problem is important, because an adaptive equalizer designed for an incorrect channel length is suboptimal: it often estimates an unnecessarily large number of parameters. Some solutions to this problem exist (e.g., attempting to estimate the "channel length", and then switching between different equalizers); however, these are suboptimal owing to the difficulty of correctly identifying the channel length, and the risk associated with an incorrect estimation of this length. Indeed, to determine the channel length is effectively a model order selection problem, for which no optimal solution is known.

We propose a novel, systematic approach to the problem under study, which circumvents the estimation of the channel length. The key idea is to model the channel impulse response via a mixture Gaussian model, which has one component for each possible channel length. The parameters of the mixture model are estimated from a received pilot sequence. We derive the optimal receiver associated with this mixture model, along with some computationally efficient approximations of it. Via numerical simulations, we show that our new method can outperform conventional adaptive Viterbi equalizers that use a fixed or an estimated channel length.

6.1.1 Problem Formulation and Background

This chapter is concerned with single-carrier transmission over a frequency-selective channel. Suppose that one transmits a sequence of complex symbols

$\{s(n)\}_{n=-N_{\rm pre}}^{N+N_{\rm post}-1}$, composed of an $N_{\rm pre}$ symbols long preamble $\{s(n)\}_{n=-N_{\rm pre}}^{-1}$ (which is known to the receiver), an N-symbol data sequence $\{s(n)\}_{n=0}^{N-1}$ (unknown to the receiver), and an $N_{\rm post}$ symbols long postamble $\{s(n)\}_{n=N}^{N+N_{\rm post}-1}$ (known to the receiver) in a standard fashion.[1] The received data then take on the form

$$y(n) = \sum_{l=0}^{L} h_l s(n-l) + e(n), \quad n = 0, \ldots, N+L-1 \tag{6.1}$$

where $\{h_l\}_{l=0}^{L}$ is the symbol-sampled baseband impulse response of the channel, and L is its length.[2] Also, in (6.1), $e(n)$ is a noise term. Note that the samples received before $n=0$ and after $n=N+L-1$ do not depend on the data symbols, so if $\{h_l\}$ is known and $e(n)$ is white, then $\{y(n)\}_{n=0}^{N+L-1}$ is a sufficient statistic for the detection of $\{s(n)\}_{n=0}^{N-1}$. If the channel $\{h_l\}$ and the properties of $e(n)$ are known, then the transmitted data $\{s(n)\}_{n=0}^{N-1}$ can be recovered from $\{y(n)\}_{n=0}^{N+L-1}$ by using standard equalization techniques [59, 60]. One of the most well known equalizers is the maximum-likelihood sequence detector (MLSD), which can be efficiently implemented via the Viterbi algorithm (VA) [61, 62].[3] If the channel $\{h_l\}_{l=0}^{L}$ is unknown to the receiver, then an estimate of it, $\{\hat{h}_l\}_{l=0}^{L}$, is usually first formed by using received pilot data, resulting in an *adaptive* equalization scheme. Such adaptive equalization is implemented in a variety of contemporary systems (the global system for mobile communications (GSM), for example [63]).

Adaptive equalization relies on the estimation of $\{h_l\}_{l=0}^{L}$, and it therefore requires L to be known. However, L may be unknown in practice, and also it may vary with the propagation environment. For example, the equalizer in a GSM phone typically assumes that $L=4$, allowing for delay spreads of up to 20μs, which may arise in hilly terrain environments [64]. However, in other environments (urban, for instance) one usually encounters much smaller delay spreads and consequently $L<4$. In such cases, an adaptive equalizer based on the assumption that $L=4$ is suboptimal because it tries to estimate a larger number of parameters than what is necessary to uniquely parameterize the channel impulse response.[4] In propagation environments with lower delay spreads, using a smaller L (i.e., assuming a shorter impulse response) would result in a smaller number of parameters to estimate, which

[1] The purpose of transmitting the known preambles and postambles typically is to facilitate equalization, and in some cases to provide the necessary means for extracting multipath diversity. In some systems, the data blocks are sequentially transmitted so that the postamble of one block overlaps with the preamble of the next.

[2] In practice, L is not finite because the channel typically is bandlimited. However, it is common practice to truncate the channel at its effective length.

[3] Unless otherwise is clear from the context, whenever we simply say "equalization" or "adaptive equalization" we refer to (adaptive) equalization using MLSD.

[4] Using the terminology in the system identification literature [1], we would say that the model used by an adaptive equalizer assuming $L=4$ is not *parsimonious*.

in turn would reduce the total channel estimation error and hence improve the equalizer performance.

6.1.2 Previous and Related Work

The possible benefit of adapting the number of taps in the equalizer to the propagation environment has been recognized before. For example, [65] suggested to switch between a receiver *with* an equalizer and a receiver *without* an equalizer, based on the estimated channel impulse response. Even such a somewhat ad hoc modification can provide a non-negligible improvement of the receiver sensitivity. Another approach, suggested in [66], is to find the number of taps by applying F-tests on a full channel model estimated using frequency domain approaches. Yet, to our knowledge, the problem of adaptive equalization with unknown channel lengths has not been systematically studied. We should note, however, that the related problem of adaptive filtering and prediction with unknown model orders (which is fundamentally different in nature as it neglects the finite-alphabet properties of a communications signal) has been studied for several decades [57, 67, 68] and also more recently reviewed in [69, 70].

One could conceive a receiver that would first try to determine an estimate of L, say $\hat{L}$, and then estimate a channel impulse response of length $\hat{L}+1$. The estimation of L would then essentially be a model order selection problem. Several order selection methods exist; e.g., see the books [2, 1, 3, 4, 5, 6] and the review papers [46, 71]. Some methods, for example the Akaike information criterion (AIC) [24], are based on minimization of the Kullback-Leibler distance [23] between the observed data and the model. Another approach is to use information-theoretic arguments to determine the length of the shortest binary sequence that can describe the data, leading to the minimum description length (MDL) criterion [35]. A third avenue is to use a Bayesian approach, leading to the Bayesian information criterion (BIC) [33, 34]. BIC is sometimes argued to be preferable as it yields consistent order estimates: One can show that under certain assumptions the probability of selecting the true model approaches 1 as the number of samples $N \to \infty$.[5] Unfortunately, it is very difficult to objectively compare the performance of model order selection methods, or even to define what one means by "optimal". This is so because there are no objective measures such as "unbiasedness" or "mean-square-error" (both of which are essential to defining optimality for continuous-parameter estimation). Furthermore, even if an "optimal" estimate $\hat{L}$ of the model order L were available, using the model associated with $\hat{L}$ for inference on the data

[5] This consistency property is *not* shared by the AIC, but in practice this is of little concern as the models being used seldom perfectly describe the highly complex reality. Furthermore, large data samples are often needed for the consistency property to become visible [5].

symbols would be suboptimal because *information is lost when one assumes that* $L = \hat{L}$: The inferences which could have been obtained when assuming other values of L are neglected [55].

6.1.3 Contributions

We present a systematic approach to the problem of equalization for channels with impulse responses of unknown length. The approach circumvents the explicit estimation of L, and therefore avoids the fundamental problems associated with methods based on model order selection combined with parameter (channel) estimation (as described above). The key idea of our approach is to parameterize the propagation channel via a Gaussian mixture model.[6] Based on the Gaussian mixture model, we derive the maximum *a posteriori* (MAP) detector, which is known to minimize the probability of a detection error. The mixture model contains a number of hyperparameters (namely, probabilities that the channel is of length $l = 0, \ldots, L$, and the associated propagation coefficients), which we estimate from a received training sequence (similarly to how the propagation channel is estimated from pilots in a conventional adaptive equalizer). We also present computationally efficient approximations to our detector. Our approach only affects the receiver design, and it is therefore fully compatible with existing transmission schemes.

Notation: Throughout the chapter, $(\cdot)^T$ is the transpose; $(\cdot)^H$ stands for the conjugate transpose; $\|\cdot\|$ stands for the Euclidean (Frobenius) norm; $|\cdot|$ denotes the determinant of a matrix; $\boldsymbol{I}$ is the identity matrix; $\Pi_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^H\boldsymbol{X})^{-1}\boldsymbol{X}^H$ is the orthogonal projector onto the column space of $\boldsymbol{X}$; and $\Pi_{\boldsymbol{X}}^{\perp} = \boldsymbol{I} - \Pi_{\boldsymbol{X}}$ is the corresponding projector onto the orthogonal complement of the column space of $\boldsymbol{X}$.

6.2 Preliminaries

To derive the optimal detector it is convenient to first establish a matrix-algebraic framework for channels with intersymbol interference. Towards this end, note that (6.1) can be written as

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{h} + \boldsymbol{e} \tag{6.2}$$

[6] It is noteworthy that estimation and detection methods based on mixture models are quite common in the statistical literature [55, 54], and in the literature on pattern recognition and artificial intelligence [72]. However, such models appear to have been much less frequently used for applications in communication theory. See, e.g., [73] for a general discussion of mixture models, in particular Gaussian mixtures.

where

$$\begin{aligned}
\boldsymbol{y} &= \begin{bmatrix} y(0) & \cdots & y(N+L-1) \end{bmatrix}^T \\
\boldsymbol{e} &= \begin{bmatrix} e(0) & \cdots & e(N+L-1) \end{bmatrix}^T \\
\boldsymbol{h} &= \begin{bmatrix} h_0 & \cdots & h_L \end{bmatrix}^T
\end{aligned} \tag{6.3}$$

and $\boldsymbol{S}$ is the following Toeplitz matrix:

$$\boldsymbol{S} = \begin{bmatrix}
s(0) & s(-1) & \cdots & s(-L) \\
s(1) & s(0) & \ddots & s(1-L) \\
\vdots & \vdots & & \vdots \\
s(L) & s(L-1) & \cdots & s(0) \\
\vdots & \vdots & & \vdots \\
s(N+L-1) & s(N+L-2) & \cdots & s(N-1)
\end{bmatrix} \tag{6.4}$$

We assume throughout that the noise $e(n)$ is white, zero-mean circularly symmetric Gaussian with variance σ^2; then $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I})$. We also assume that $\boldsymbol{s}$ and $\boldsymbol{e}$ are independent, where we introduce the notation $\boldsymbol{s} = \begin{bmatrix} s(0) & \cdots & s(N-1) \end{bmatrix}^T$. Finally, we let $\mathcal{S}$ be the set of all possible vectors $\boldsymbol{s}$.

The detector that minimizes the probability of an incorrectly detected sequence $\boldsymbol{s}$ is the maximum *a posteriori* detector, i.e., the detector that maximizes $P(\boldsymbol{s}|\boldsymbol{y})$ [74]. If all symbol sequences $\{\boldsymbol{s}\}$ are equiprobable, then $\hat{\mathbf{s}}_{\mathrm{MAP}} = \operatorname{argmax}_{\boldsymbol{s} \in \mathcal{S}} P(\boldsymbol{s}|\boldsymbol{y}) = \operatorname{argmax}_{\boldsymbol{s} \in \mathcal{S}} p(\boldsymbol{y}|\boldsymbol{s})$, where $p(\boldsymbol{y}|\boldsymbol{s})$ is the likelihood function of the received data. If $\boldsymbol{h}$ were known to the receiver, then under the assumptions made, optimal detection of $\boldsymbol{s}$ would amount to minimizing the Euclidean distance

$$\|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{h}\|^2 = \sum_{n=0}^{N+L-1} \left| y(n) - \sum_{l=0}^{L} h_l s(n-l) \right|^2 \tag{6.5}$$

which can be accomplished by using the Viterbi algorithm. Inserting a suitable estimate $\hat{\boldsymbol{h}}$ of $\boldsymbol{h}$ (obtained, for example, via maximum-likelihood) in (6.5) yields a conventional adaptive MLSD equalizer. While such a receiver is in no way optimal, not even when $\hat{\boldsymbol{h}}$ is an "optimal" estimate of $\boldsymbol{h}$, it usually works in a satisfactory manner and it has become common practice (see, e.g., Chapter 11 of [59] for a general discussion of adaptive equalizers).

6.3 Gaussian Mixture Channel Model

The assumption that underlies our new approach is that the propagation channel $\boldsymbol{h}$ can be parameterized via a Gaussian mixture model that has one component $\mathcal{M}_n$ (with the associated probability $P(\mathcal{M}_n)$) for each possible channel length $n = 0, \ldots, L$:

$$\begin{array}{llll} \mathcal{M}_0: & h_0 & \sim N(\mu_0, \Lambda_0); & h_1 = \cdots = h_L = 0 \\ \mathcal{M}_1: & [h_0 \; h_1]^T & \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1); & h_2 = \cdots = h_L = 0 \\ \quad \vdots & & & \\ \mathcal{M}_L: & [h_0 \; \cdots \; h_L]^T & \sim N(\boldsymbol{\mu}_L, \boldsymbol{\Lambda}_L) & \end{array} \tag{6.6}$$

(Hereafter we let L denote the *maximum possible* channel length; hence, there are $L+1$ possible models.) To interpret the mixture model, one can think that the mechanism which generates $\boldsymbol{h}$ first chooses the channel length n (with probability $P(\mathcal{M}_n)$) and then generates a random channel $\boldsymbol{h}$ that has the statistical distribution associated with $\mathcal{M}_n$. Specifically, with probability $P(\mathcal{M}_0)$, the channel is frequency-flat; with probability $P(\mathcal{M}_1)$, the channel has one fading tap; and with probability $P(\mathcal{M}_n)$ it has $n+1$ taps. Note that the total pdf of $\boldsymbol{h}$ is a Gaussian mixture density. Of course, if one knows *a priori* that the channel length is contained in a given interval, one can restrict the number of possible models to a subset of those in (6.6). In this chapter we will assume for simplicity that all models in (6.6) are permissible; the extension to the general case is immediate but it does require some additional notation.

We wish to stress that although at this point it appears that our model would be valid only for channels that actually are Gaussian (or mixture Gaussian), we later (see Section 6.4.1) show that if no *a priori* information on the true channel is available at all, then conditioned on a received training sequence the true channel will have a Gaussian density. Therefore, the proposed mixture Gaussian model is extremely general and it does *not* require the true channel to have a mixture Gaussian density *a priori*.

In practice, $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n)\}$ are unknown. However, as a preparation, in the next section we first assume that we know these quantities and derive the optimal detector under this assumption. Later, in Section 6.4, we present a method for estimating $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n)\}$ from a received training (pilot) sequence, thereby obtaining what we call a "trained multi-model detector."

6.3.1 The Optimal Receiver for Known $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n), \sigma^2\}$

We next derive the maximum *a posteriori* (MAP) detector for the mixture model, assuming that $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n), \sigma^2\}$ are *known*. We have that

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{s}) &= \sum_{n=0}^{L} P(\mathcal{M}_n) p(\boldsymbol{y}|\boldsymbol{s}, \mathcal{M}_n) \\
&= \sum_{n=0}^{L} P(\mathcal{M}_n) \int_{\mathbb{C}^n} p(\boldsymbol{y}|\boldsymbol{h}_n, \boldsymbol{s}) p(\boldsymbol{h}_n) d\boldsymbol{h}_n \\
&= \sum_{n=0}^{L} P(\mathcal{M}_n) \int_{\mathbb{C}^n} \frac{1}{(\pi\sigma^2)^{N+L}} \frac{1}{\pi^{n+1}|\boldsymbol{\Lambda}_n|} \\
&\quad \cdot \exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{S}_n\boldsymbol{h}_n\|^2 - (\boldsymbol{h}_n - \boldsymbol{\mu}_n)^H \boldsymbol{\Lambda}_n^{-1}(\boldsymbol{h}_n - \boldsymbol{\mu}_n)\right) d\boldsymbol{h}_n
\end{aligned}
\tag{6.7}
$$

where $\boldsymbol{S}_n$ is the matrix made from the first $n+1$ columns of the matrix $\boldsymbol{S}$ in (6.4), and $\boldsymbol{h}_n = \begin{bmatrix} h_0 & \cdots & h_n \end{bmatrix}^T$. The integral in (6.7) can be evaluated by some straightforward but somewhat tedious algebra. Note first that

$$
\|\boldsymbol{y} - \boldsymbol{S}_n\boldsymbol{h}_n\|^2 = (\boldsymbol{h}_n - \tilde{\boldsymbol{h}}_n)^H \boldsymbol{S}_n^H \boldsymbol{S}_n (\boldsymbol{h}_n - \tilde{\boldsymbol{h}}_n) + \boldsymbol{y}^H \Pi_{\boldsymbol{S}_n}^{\perp} \boldsymbol{y} \tag{6.8}
$$

where $\tilde{\boldsymbol{h}}_n = (\boldsymbol{S}_n^H\boldsymbol{S}_n)^{-1}\boldsymbol{S}_n^H\boldsymbol{y}$. It follows that

$$
\begin{aligned}
&\|\boldsymbol{y} - \boldsymbol{S}_n\boldsymbol{h}_n\|^2 + \sigma^2(\boldsymbol{h}_n - \boldsymbol{\mu}_n)^H\boldsymbol{\Lambda}_n^{-1}(\boldsymbol{h}_n - \boldsymbol{\mu}_n) \\
=&(\boldsymbol{h}_n - \tilde{\boldsymbol{h}}_n)^H(\boldsymbol{S}_n^H\boldsymbol{S}_n)(\boldsymbol{h}_n - \tilde{\boldsymbol{h}}_n) + \sigma^2(\boldsymbol{h}_n - \boldsymbol{\mu}_n)^H\boldsymbol{\Lambda}_n^{-1}(\boldsymbol{h}_n - \boldsymbol{\mu}_n) + \boldsymbol{y}^H\Pi_{\boldsymbol{S}_n}^{\perp}\boldsymbol{y} \\
=&(\boldsymbol{h}_n - \bar{\boldsymbol{h}}_n)^H(\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H\boldsymbol{S}_n)(\boldsymbol{h}_n - \bar{\boldsymbol{h}}_n) + \Delta + \boldsymbol{y}^H\Pi_{\boldsymbol{S}_n}^{\perp}\boldsymbol{y}
\end{aligned}
\tag{6.9}
$$

where we defined

$$
\bar{\boldsymbol{h}}_n =(\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H\boldsymbol{S}_n)^{-1}(\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H\boldsymbol{S}_n\tilde{\boldsymbol{h}}_n) \tag{6.10}
$$

and

$$
\begin{aligned}
\Delta =&\tilde{\boldsymbol{h}}_n^H\boldsymbol{S}_n^H\boldsymbol{S}_n\tilde{\boldsymbol{h}}_n + \sigma^2\boldsymbol{\mu}_n^H\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n \\
&- (\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H\boldsymbol{S}_n\tilde{\boldsymbol{h}}_n)^H(\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H\boldsymbol{S}_n)^{-1}(\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H\boldsymbol{S}_n\tilde{\boldsymbol{h}}_n) \\
=&\boldsymbol{y}^H\Pi_{\boldsymbol{S}_n}\boldsymbol{y} + \sigma^2\boldsymbol{\mu}_n^H\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n \\
&- (\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H\boldsymbol{y})^H(\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H\boldsymbol{S}_n)^{-1}(\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H\boldsymbol{y})
\end{aligned}
\tag{6.11}
$$

Consequently,

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{s}) = & \frac{1}{(\pi\sigma^2)^{N+L}} \sum_{n=0}^{L} P(\mathcal{M}_n) \frac{\sigma^{2(n+1)}|\boldsymbol{\Lambda}_n^{-1}|}{|\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n|} e^{-\frac{\Delta + \boldsymbol{y}^H \Pi^{\perp}_{\boldsymbol{S}_n} \boldsymbol{y}}{\sigma^2}} \\
& \cdot \underbrace{\int_{\mathbb{C}^n} \frac{|\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n|}{(\pi\sigma^2)^{n+1}} e^{-\frac{1}{\sigma^2}(\boldsymbol{h}_n - \bar{\boldsymbol{h}}_n)^H (\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n)(\boldsymbol{h}_n - \bar{\boldsymbol{h}}_n)} d\boldsymbol{h}_n}_{=1} \\
= & \frac{1}{(\pi\sigma^2)^{N+L}} \sum_{n=0}^{L} P(\mathcal{M}_n) \frac{\sigma^{2(n+1)}|\boldsymbol{\Lambda}_n^{-1}|}{|\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n|} \\
& \cdot \exp\Bigg(-\frac{\|\boldsymbol{y}\|^2 + \sigma^2 \boldsymbol{\mu}_n^H \boldsymbol{\Lambda}_n^{-1} \boldsymbol{\mu}_n}{\sigma^2} \\
& - \frac{(\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H \boldsymbol{y})^H (\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n)^{-1} (\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H \boldsymbol{y})}{\sigma^2} \Bigg)
\end{aligned} \tag{6.12}$$

(The integrand in (6.12) is the probability density function of a Gaussian random variable, which integrates to unity over $\mathbb{C}^n$.) In conclusion, we have that

$$\begin{aligned}
\hat{\boldsymbol{s}}_{\mathrm{MAP}} = & \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} \Bigg\{ \sum_{n=0}^{L} P(\mathcal{M}_n) \frac{\sigma^{2(n+1)}|\boldsymbol{\Lambda}_n^{-1}|}{|\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n|} \\
& \cdot e^{-\frac{\sigma^2 \boldsymbol{\mu}_n^H \boldsymbol{\Lambda}_n^{-1} \boldsymbol{\mu}_n - (\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H \boldsymbol{y})^H (\sigma^2\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{S}_n^H \boldsymbol{S}_n)^{-1} (\sigma^2\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{S}_n^H \boldsymbol{y})}{\sigma^2}} \Bigg\}
\end{aligned} \tag{6.13}$$

6.3.2 Discussion of the Optimal Receiver

The detection rule (6.13) is optimal (under the assumed model) in the sense that it minimizes the probability that the detected sequence differs from the transmitted one. However, in general one does not know the *a priori* probabilities $\{P(\mathcal{M}_n)\}$, nor the quantities $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$. Since (6.13) is a Bayesian detector, one could think of treating $\{P(\mathcal{M}_n), \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$ as hyperparameters and concentrate them out by maximizing (6.12) with respect to $\{P(\mathcal{M}_n), \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$. Although this makes sense intuitively, and it would result in a "type-II maximum-likelihood" detector [75], we do not pursue this approach here. First, maximizing (6.12) with respect to $\{P(\mathcal{M}_n), \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$ (and $\boldsymbol{s}$) appears to be an intractable problem. Second, studies of related problems [76] have indicated that concentrating out hyperparameters from likelihood functions used for symbol detection can lead to certain technical difficulties. Our trained multi-model detector, presented next, provides a way of estimating $\{P(\mathcal{M}_n), \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$ from a received training sequence, thereby obviating the problems associated with the direct use of (6.13).

6.4 Training-Based Estimation of the Mixture Model Parameters

We next describe how the parameters (viz., $\{P(\mathcal{M}_n), \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$) in the multi-model detector (6.13) can be estimated from a training sequence. Specifically, we will show that if the noise received during the training sequence is Gaussian, and a linear estimator is used, then the estimates $\{\hat{\boldsymbol{h}}_n\}$ of $\{\boldsymbol{h}_n\}$ are Gaussian too and the covariance matrices $\{\boldsymbol{\Lambda}_n\}$ are easily obtained. If $\hat{\boldsymbol{h}}_n$ is Gaussian conditioned on $\boldsymbol{h}_n$, then we show that $\boldsymbol{h}_n$ is Gaussian conditioned on the training data, and hence the Gaussian mixture model assumption made in (6.6) is satisfied. Finally, we use a theorem in [33] to estimate $\{P(\mathcal{M}_n)\}$.

6.4.1 Estimation of $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$ from a Training Sequence

Suppose we transmit a training sequence $s_t(n)$ and use it to estimate the channel at the receiver. The received data during the training sequence, of length N_t, can be written

$$\boldsymbol{y}_t = \boldsymbol{S}_t \boldsymbol{h} + \boldsymbol{e}_t \tag{6.14}$$

where $\boldsymbol{y}_t$ and $\boldsymbol{e}_t$ are defined in an obvious way and where $\boldsymbol{S}_t$ is the matrix in (6.4) associated with $s_t(n)$. For a given model $\mathcal{M}_n$, and assuming Gaussian noise, it is well known that under the assumptions made above, the ML estimate of $\boldsymbol{h}_n$ is given by

$$\hat{\boldsymbol{h}}_n = \left(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn}\right)^{-1} \boldsymbol{S}_{tn}^H \boldsymbol{y}_t \tag{6.15}$$

where $\boldsymbol{S}_{tn}$ is the matrix made from the first n columns of $\boldsymbol{S}_t$. Since $\hat{\boldsymbol{h}}_n = \boldsymbol{h}_n + \left(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn}\right)^{-1} \boldsymbol{S}_{tn}^H \boldsymbol{e}_t$, it directly follows that

$$\hat{\boldsymbol{h}}_n \sim N\Big(\boldsymbol{h}_n, \sigma^2 \left(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn}\right)^{-1}\Big) \quad \text{conditioned on } \boldsymbol{h}_n \tag{6.16}$$

To use the mixture model detector, we need to find the density of $\boldsymbol{h}$ conditioned on $\hat{\boldsymbol{h}}$ (i.e., we need to “invert” the conditional density $p(\hat{\boldsymbol{h}}|\boldsymbol{h})$ in (6.16)). Towards this end, let us assume that $\boldsymbol{h}$ is a random variable of which we have no prior knowledge (before observing the training). Let us model the *a priori* density of this random variable via $\boldsymbol{h} \sim N(\boldsymbol{0}, \gamma \boldsymbol{I})$, where γ is a very large number. Under this assumption, $\boldsymbol{h}$ and $\hat{\boldsymbol{h}}$ are jointly Gaussian with zero mean and covariance matrix

$$\mathcal{E}\left\{\begin{bmatrix}\boldsymbol{h}\\ \hat{\boldsymbol{h}}\end{bmatrix}\begin{bmatrix}\boldsymbol{h}^H & \hat{\boldsymbol{h}}^H\end{bmatrix}\right\} = \begin{bmatrix}\gamma\boldsymbol{I} & \gamma\boldsymbol{I}\\ \gamma\boldsymbol{I} & \gamma\boldsymbol{I} + \sigma^2(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn})^{-1}\end{bmatrix} \tag{6.17}$$

It follows that conditioned on $\hat{\boldsymbol{h}}$, $\boldsymbol{h}$ is Gaussian with mean $\gamma\boldsymbol{I}(\gamma\boldsymbol{I} + \sigma^2(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn})^{-1})^{-1}\hat{\boldsymbol{h}}$ and covariance matrix $\gamma\boldsymbol{I} - \gamma\boldsymbol{I}(\gamma\boldsymbol{I} + \sigma^2(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn})^{-1})^{-1}\gamma\boldsymbol{I}$.

In the limit when $\gamma \to \infty$ (corresponding to the case of no *a priori* information on $\boldsymbol{h}$), we find that

$$\lim_{\gamma\to\infty} \operatorname{cov}(\boldsymbol{h}|\hat{\boldsymbol{h}}) = \lim_{\gamma\to\infty} \left(\gamma \boldsymbol{I} - \gamma \boldsymbol{I}(\gamma \boldsymbol{I} + \sigma^2(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn})^{-1})^{-1}\gamma \boldsymbol{I}\right) = \sigma^2(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn})^{-1} \tag{6.18}$$

This means that asymptotically as $\gamma \to \infty$, the density of $\boldsymbol{h}$ conditioned on $\hat{\boldsymbol{h}}$ is formally equal to that of $\hat{\boldsymbol{h}}$ conditioned on $\boldsymbol{h}$:

$$\boldsymbol{h}_n \sim N\left(\hat{\boldsymbol{h}}_n, \sigma^2\left(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn}\right)^{-1}\right) \quad \text{conditioned on } \boldsymbol{y}_t \text{ and } \boldsymbol{S}_{tn} \tag{6.19}$$

This result means that the true channel is Gaussian conditioned on the received training, and therefore that the assumptions of the mixture model are satisfied. Note that for a sufficiently large number of pilots, the central limit theorem implies that $\boldsymbol{h}$ and $\hat{\boldsymbol{h}}$ are approximately Gaussian even in the event that the noise $\boldsymbol{e}_t$ is non-Gaussian.

6.4.2 Estimation of $\{P(\mathcal{M}_n)\}$

To estimate the probabilities $\{P(\mathcal{M}_n)\}$, we use a result from [33] (also presented in a more popularized form in [46, 69]). By assuming that $p(\boldsymbol{h}_n)$ has a dominant peak at $\hat{\boldsymbol{h}}_n$, one can show that

$$\log p(\boldsymbol{y}_t|\mathcal{M}_n) \approx -\frac{1}{2}\mathrm{BIC}_n + \text{const.} \tag{6.20}$$

where

$$\mathrm{BIC}_n \triangleq -2\log[p(\boldsymbol{y}_t|\mathcal{M}_n)]_{\boldsymbol{h}_n:=\hat{\boldsymbol{h}}_n} + (2(n+1)+1)\log 2N_t \tag{6.21}$$

and where the approximation error in (6.20) goes to zero when the amount of training increases without bound. In (6.21), $[p(\boldsymbol{y}_t|\mathcal{M}_n)]_{\boldsymbol{h}_n:=\hat{\boldsymbol{h}}_n}$ is the likelihood function of the data with $\boldsymbol{h}_n$ replaced by $\hat{\boldsymbol{h}}_n$, and σ^2 replaced by $\hat{\sigma}_n^2$; that is,

$$[p(\boldsymbol{y}_t|\mathcal{M}_n)]_{\boldsymbol{h}_n:=\hat{\boldsymbol{h}}_n} = \text{const.} \cdot \frac{1}{(\hat{\sigma}_n^2)^{N_t}} \tag{6.22}$$

where

$$\hat{\sigma}_n^2 = \frac{1}{N_t}\left\|\boldsymbol{y}_t - \boldsymbol{S}_{tn}\hat{\boldsymbol{h}}_n\right\|^2 = \frac{1}{N_t}\left\|\Pi_{\boldsymbol{S}_{tn}}^{\perp}\boldsymbol{y}_t\right\|^2 \tag{6.23}$$

Hence, to within an additive constant,

$$\mathrm{BIC}_n = 2N_t \log\left\|\Pi_{\boldsymbol{S}_{tn}}^{\perp}\boldsymbol{y}_t\right\|^2 + (2(n+1)+1)\log 2N_t \tag{6.24}$$

It is clear that

$$P(\mathcal{M}_n|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|\mathcal{M}_n)P(\mathcal{M}_n)}{p(\boldsymbol{y}_t)} \tag{6.25}$$

Assuming that all models (channel lengths) are equally probable *a priori* (before observing the training), i.e., $P(\mathcal{M}_n) = 1/(L+1)$, and using the fact that $\sum_{n=0}^{L} P(\mathcal{M}_n|\boldsymbol{y}_t) = 1$, (6.25) and (6.20) lead to the following estimate of the *a posteriori* probability $P(\mathcal{M}_n)$:

$$\hat{P}(\mathcal{M}_n|\boldsymbol{y}_t) = \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}_n\right)}{\sum_{k=0}^{L}\exp\left(-\frac{1}{2}\mathrm{BIC}_k\right)} \tag{6.26}$$

where BIC_k is given by (6.24). Note that the influence of the *a priori* assumption on $\{P(\mathcal{M}_n)\}$ (namely, that $P(\mathcal{M}_n) = 1/(L+1)$), will diminish as the amount of training data increases. Such an assumption, however, is necessary to make in order to use BIC to estimate the *a posteriori* model probabilities.

6.5 The Trained Multi-Model Detector and Its Implementation

In this section we first discuss the detector that results from using estimates of $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n), \sigma^2\}$ in (6.13). We then go on to discuss approximations of the so-obtained adaptive detector.

6.5.1 Trained Multi-Model Detector

We have found in Section 6.4.1 that conditioned on the observed training data $\boldsymbol{y}_t$, the channels $\{\boldsymbol{h}_n\}$ are Gaussian distributed according to (6.19). Thus, the assumptions of the Gaussian mixture model (6.6) are satisfied. Furthermore, given $\{\boldsymbol{y}_t, \boldsymbol{S}_t\}$, (6.19), (6.23) and (6.26) provide estimates of $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n), \sigma^2\}$. This means that we can apply the multi-model detector (6.13) with

$$\begin{aligned}
\boldsymbol{\mu}_n &:= \hat{\boldsymbol{h}}_n = \left(\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn}\right)^{-1} \boldsymbol{S}_{tn}^H \boldsymbol{y}_t \\
\boldsymbol{\Lambda}_n &:= \hat{\sigma}_n^2 (\boldsymbol{S}_{tn}^H \boldsymbol{S}_{tn})^{-1} \\
\sigma^2 &:= \hat{\sigma}_L^2 = \frac{1}{N_t}\left\|\Pi_{\boldsymbol{S}_{tL}}^{\perp} \boldsymbol{y}_t\right\|^2 \\
P(\mathcal{M}_n) &:= \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}_n\right)}{\sum_{k=0}^{L}\exp\left(-\frac{1}{2}\mathrm{BIC}_k\right)}
\end{aligned} \tag{6.27}$$

where BIC_k is given by (6.24). Note that the estimates in (6.27) are *a posteriori* with respect to the training data but *a priori* with respect to the unknown symbol data sample.

In (6.27), we use the noise variance estimate associated with the maximum channel order, viz. $\hat{\sigma}_L^2$. We choose to do so because $\hat{\sigma}_L^2$ is usually a better estimate of σ^2 than $\hat{\sigma}_n^2$ is. Indeed, when n is less than the true channel length, then the more "natural" estimate $\hat{\sigma}_n^2$ tends to be an overestimate of σ^2. This is so because in this case, $\hat{\sigma}_n^2$ will be the sum of the noise variance and a component associated with the residual modeling error caused by the use of an insufficient number of parameters.

The detector resulting from (6.13) and (6.27) can be explicitly spelled out (by inserting (6.27) into (6.13)), but in the interest of brevity we avoid presenting the resulting, somewhat messy expression. We stress that the so-obtained detector is not fully optimal, despite the optimality of (6.13) for known $\{\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, P(\mathcal{M}_n), \sigma^2\}$; indeed, the procedure of first extracting information from the training and then using that information in the detector (6.13) is suboptimal in itself. Nevertheless, it is noteworthy that the detector (6.13)+(6.27) extracts more information (viz., $\{\boldsymbol{\mu}_n\}$, $\{\boldsymbol{\Lambda}_n\}$, σ^2 and $\{P(\mathcal{M}_n)\}$) from the training than what a conventional training-based detector typically does; indeed, the latter typically only determines a fixed-order channel estimate $\hat{\boldsymbol{h}}$ and uses it in the detector as if it were equal to the true channel. Such a conventional training-based detector is a special case that we can obtain by successively simplifying the equations (6.13) and (6.27), as explained in Section 6.5.2.

As a further remark, note that when we derived the optimal MAP detector in (6.13), we assumed that σ^2 was a deterministic constant. However, the estimate of σ^2, viz. $\hat{\sigma}^2$, obtained from the training is a random variable. One could therefore think of obtaining the distribution of σ^2 conditioned on the training data, and appropriately modify the detector (6.13) to optimally make use of this information as well. However, we do not pursue this path. First, since the noise variance can usually be fairly accurately estimated, and since symbol detection criteria do not seem to be excessively sensitive to the estimation of noise powers in general, we expect the gain obtained by incorporating the distribution of σ^2 (conditioned on the training data) into (6.13) to be small. Second, it appears to be difficult to obtain the associated MAP detector in closed form.

6.5.2 Possible Approximations of the Trained Multi-Model Detector (6.13)+(6.27)

The problem of directly solving (6.13), using (6.27), is computationally rather hard. One can obtain simplified forms of this problem by performing a series of approximations as described next.

Approximation 1: $\boldsymbol{h}_n = \hat{\boldsymbol{h}}_n$: The first possible approximation is to assume that $\boldsymbol{h}_n$ is perfectly estimated from the training. In loose words, this amounts to assuming that $\boldsymbol{\mu}_n = \hat{\boldsymbol{h}}_n$ and $\boldsymbol{\Lambda}_n = \mathbf{0}$, whereby the distribution of $\hat{\boldsymbol{h}}_n$ becomes a Dirac-impulse at $\boldsymbol{h}_n = \hat{\boldsymbol{h}}_n$. In this case, the integral over $\boldsymbol{h}_n$ disappears from (6.7), and we are left with:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{s}) &= \sum_{n=0}^{L} P(\mathcal{M}_n) p(\boldsymbol{y}|\hat{\boldsymbol{h}}_n, \boldsymbol{s}) \\ &= \frac{1}{(\pi\sigma^2)^{N+L}} \sum_{n=0}^{L} P(\mathcal{M}_n) \exp\left(-\frac{1}{\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{S}_n\hat{\boldsymbol{h}}_n\right\|^2\right) \end{aligned} \tag{6.28}$$

Approximation 2: Weighted LS metric: Since the computational complexity of (6.28) is prohibitively high for long sequences, one can also think of the following further approximation,

$$\begin{aligned} \hat{\boldsymbol{s}} &= \underset{\boldsymbol{s}}{\operatorname{argmin}} \left\{ \sum_{n=0}^{L} P(\mathcal{M}_n) \left\|\boldsymbol{y} - \boldsymbol{S}_n\hat{\boldsymbol{h}}_n\right\|^2 \right\} \\ &= \underset{\boldsymbol{s}}{\operatorname{argmin}} \left\{ \sum_{k=0}^{N+L-1} \left(\sum_{n=0}^{L} P(\mathcal{M}_n) \left| y(k) - \sum_{l=0}^{n} \hat{h}_l^n s(k-l) \right|^2 \right) \right\} \end{aligned} \tag{6.29}$$

which is easily implemented by a standard VA using a vector-valued branch metric. The approximation involved in going from (6.28) to (6.29) is rather *ad hoc*, and indeed only motivated by the fact that (6.29) resembles an order-averaged least-squares metric.

Approximation 3: $\mathcal{M}_n = \mathcal{M}_{\hat{n}}$: As another further possible approximation of (6.28), one could assume that the model $\mathcal{M}_n$ with the highest *a posteriori* probability $P(\mathcal{M}_n)$ given by (6.26) is the true one, thereby discarding most of the information in $\{\hat{P}(\mathcal{M}_n)\}$. This would correspond to first estimating the channel length and the associated impulse response, and then using these estimated quantities in lieu of the unknown ones. In this case, the sum over n in (6.7) disappears, and one is left with the decision metric

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}\in\mathcal{S}}{\operatorname{argmin}} \left\|\boldsymbol{y} - \boldsymbol{S}_{\hat{n}}\hat{\boldsymbol{h}}_{\hat{n}}\right\|^2 = \underset{\boldsymbol{s}}{\operatorname{argmin}} \left\{ \sum_{k=0}^{N+L-1} \left| y(k) - \sum_{l=0}^{\hat{n}} \hat{h}_l^{\hat{n}} s(k-l) \right|^2 \right\} \tag{6.30}$$

where

$$\hat{n} = \underset{n\in\{0,\ldots,L\}}{\operatorname{argmax}} \hat{P}(\mathcal{M}_n|\boldsymbol{y}_t) = \underset{n}{\operatorname{argmin}} \mathrm{BIC}_n \tag{6.31}$$

Equation (6.30) is the conventional MLSD associated with the most likely model and the solution can be easily obtained with a standard VA. Soft-decisions could also be obtained by using a soft-output Viterbi algorithm (SOVA) [77] on (6.30). For more details on implementing a SOVA for this problem, see [78, 79].

6.6 Numerical Examples

We illustrate the error performance of the previously introduced detectors via Monte-Carlo simulation. We use a simple channel model to compare the performances of the trained multi-model detector (see Section 6.5.1) and the possible approximations of it (see Section 6.5.2).

Example 1 (performance of trained multiple model detector for short simple channel and short data sequence): We use a short data sequence of length $N = 6$ and a channel with $L + 1$ different exponentially decaying delay profiles, all occurring with the same probability $1/(L + 1)$. For the lth delay profile ($l \in \{1, 2, \ldots, L + 1\}$):

$$\mathcal{E}\left\{\left[|h_0|^2\ |h_1|^2\ \cdots\ |h_{L-1}|^2\right]\right\} = \frac{[\ 1\ e^{-0.25}\ \cdots\ e^{-0.25(l-1)}\ 0\ \cdots\ 0]}{\sum_{k=0}^{l-1} e^{-0.25k}} \tag{6.32}$$

where the scaling factor in the denominator ensures that all delay profiles have unit energy. In this example we have set $L = 2$; hence the channel has up to $L + 1 = 3$ non-zero taps. A 4 samples long sequence, $\mathbf{s}_t = [-1, -1, 1, -1]^T$, was used as training sequence for the estimation of $\{P(\mathcal{M}_n), \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, \sigma^2\}$ (other pilot symbol sequence lengths give similar results, but the difference between the methods tend to decrease when the length of the pilot symbol sequence increases).

Figure 6.1 shows the bit-error-rate (BER) for: (a) adaptive equalization with a 1 tap channel estimate; (b) adaptive equalization, 2 taps; (c) adaptive equalization, 3 taps; (d) adaptive equalization, 4 taps; (e) adaptive multi-model equalization using (6.13)+(6.27); and (f) coherent detection (using the true channel). Here, (a)–(d) were obtained using (6.30) with $\hat{n}$ = number of taps − 1. The improvement obtained by using the multi-model equalizer instead of the conventional adaptive MLSD (with 3 taps) is about 1 dB in this example. The performances of the conventional adaptive MLSDs using 1 or 2 taps are poor (of course), but shown for completeness. Note that the adaptive MLSD using 4 taps always uses a non-parsimonious parameterization of the channel, as it has to estimate at least one unnecessary parameter; this is why it performs worse than the MLSD using 3 taps.

In Figure 6.2, we compare the performance of conventional adaptive equalization using 3 taps, adaptive multi-model equalization (using (6.13) + (6.27)), and the various approximations in (6.28), (6.29), and (6.30). The figure shows how successive approximations of the trained multi-model equalizer degrade the performance.

Example 2 (performance of approximated multiple model detector for a long sequence): We next use a longer sequence ($N = 150$) and a channel with up to $L + 1 = 5$ taps with the exponentially decaying delay profiles in (6.32) (with $L = 4$), to compare the performance of conventional Viterbi equalization with that of the approximate multi-model detectors. A 7 samples long sequence, $\mathbf{s}_t = [-1, -1, 1, -1, -1, 1, -1]^T$, was used as training signal. Figure 6.3 shows the BER for: (a) conventional adaptive MLSD with a 5 taps channel estimate; (b) adaptive equalization using the estimated model (6.30); (c) detection using (6.29); and (d) coherent detection. The performances of the adaptive multi-model detector given by (6.13)+(6.27) and of the detector (6.28) are not shown, as these detectors are computationally prohibitive in this example. A Viterbi equalizer was used to minimize all decision metrics. In this example as well, the different approximations of the multi-model equalizer (6.13) outperform the conventional adaptive Viterbi equalizer; the improvement is about 1.5 dB for this example. The difference between the two approximations (6.29) and (6.30) (note that the latter is *not* a multi-model approach, but a single model approach using the model structure selected by BIC) is not very large in this example, but the multi-model approach of (6.29) has a slight advantage.

6.7 Concluding Remarks

We have devised mixture model-based adaptive equalizers for communication channels that can be modelled as finite-impulse-response filters with unknown length. The performances of the proposed receivers were evaluated numerically by using a simple channel model. The proposed receivers have shown promising performance. If the increased computational complexity can be afforded, they appear to be practically viable and attractive possibilities for improving the sensitivity performance of equalizers in wireless systems.

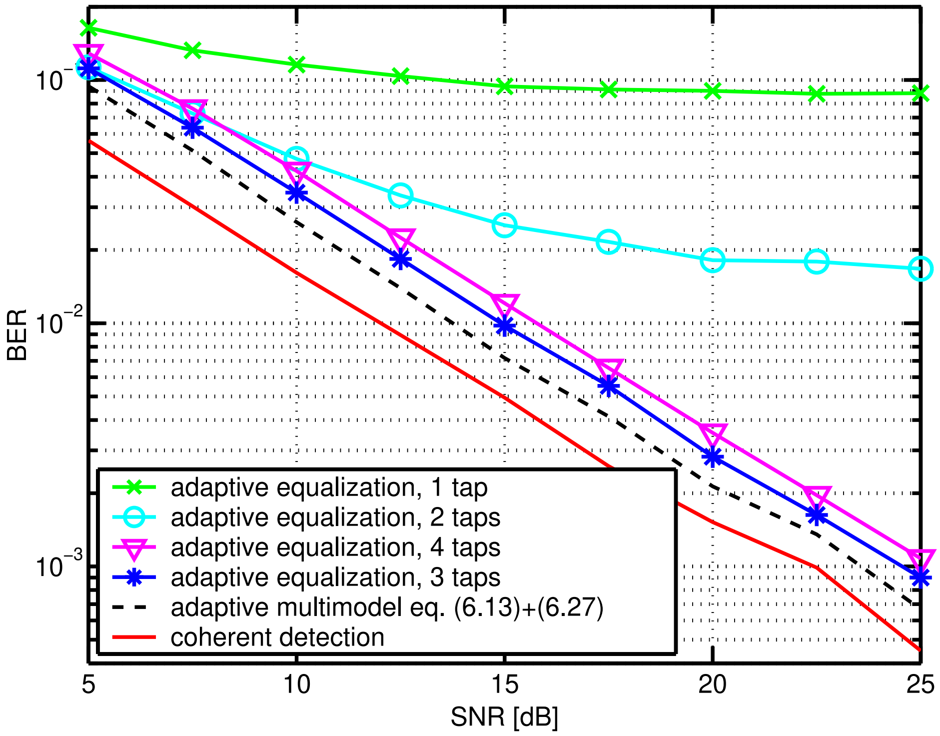


Figure 6.1: Comparison of multi-model equalization and conventional adaptive Viterbi equalization for a short sequence ($N = 6$) and a channel with at most three taps ($L = 2$). The pilot sequence had length 4.

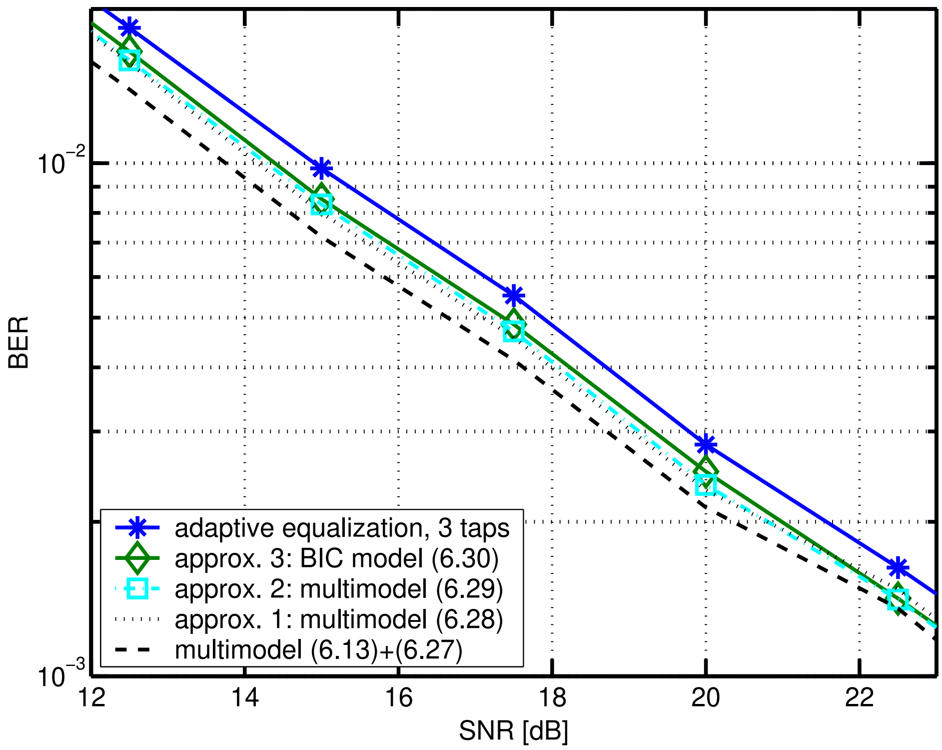


Figure 6.2: Comparison of different approximations to the adaptive multimodel equalizer for a short sequence ($N = 6$), and a channel with a maximum of three taps ($L = 2$). The pilot sequence had length 4.

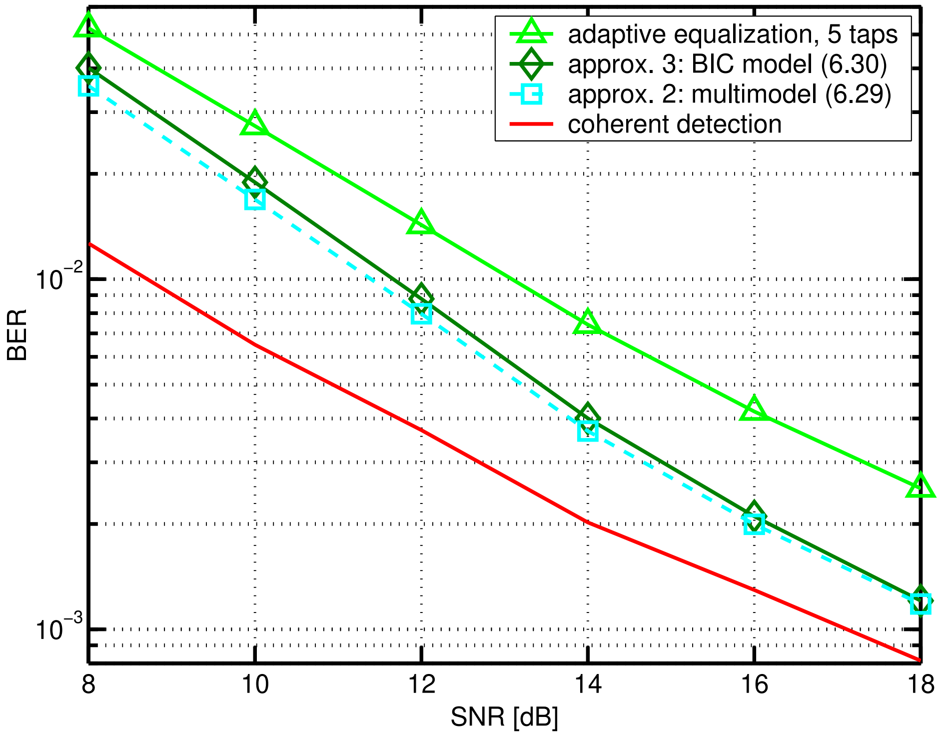


Figure 6.3: Comparison of adaptive multi-model equalization and conventional Viterbi equalization for a long sequence ($N = 150$), and a channel with up to five taps ($L = 4$). The pilot sequence had length 7.

Chapter 7

A model averaging approach for equalizing sparse communication channels

7.1 Introduction

A sparse model is a model for which many parameters equal zero. In this chapter we consider a model averaging [55] approach (multi-modeling, or channel averaging, in communication applications) based on the Information Criterion approach (IC) which consists of estimating several sparse communication channels which then are used to obtain an averaged channel for equalization.

In what follows, we will use the following definitions:

$$
\begin{aligned}
\mathbf{y} \in \mathbb{C}^N &= [y(0) \ \cdots \ y(N-1)]^T, \text{ the received data vector} \\
\mathbf{h}_L^{(n)} \in \mathbb{C}^{L+1} &= [h_0^{(n)} \ \cdots \ h_L^{(n)}]^T, \text{ the impulse response of a sparse channel} \\
&\quad \text{with a maximum order } L \text{ and } n \leq L+1 \text{ non-zero taps} \\
p_n(\mathbf{y}|\mathbf{h}_L^{(n)}) &= \text{the data likelihood associated with a channel with the} \\
&\quad \text{impulse response } \mathbf{h}_L^{(n)} \\
\hat{\mathbf{h}}_L^{(n)} &= \arg\max_{\mathbf{h}_L^{(n)}} p_n(\mathbf{y}|\mathbf{h}_L^{(n)}) = \text{the Maximum Likelihood (ML)} \\
&\quad \text{estimate of } \mathbf{h}_L^{(n)}.
\end{aligned}
$$

We are concerned with estimating the unknown channel $\mathbf{h}_L$ via the transmission of a known sequence of symbols, $\{s(t)\}_{t=-L}^{N-1}$. The data collected at the receiver take on the form

$$
y(t) = \sum_{l=0}^{L} s(t-l)h_l + e(t) = \mathbf{s}_L^T(t)\mathbf{h}_L + e(t), \quad t = 0, \ldots, N-1 \tag{7.1}
$$

where $\mathbf{s}_L^T(t) = [s(t) \ \cdots \ s(t-L)]$ and $e(t)$ denotes a noise which we assume to be zero mean, white and circularly Gaussian with variance σ^2. We also assume the unknown channel $\mathbf{h}_L$ to be *sparse*, i.e. "many" of the elements in $\mathbf{h}_L$ are zero. We further assume that the known input signal $\{s(t)\}$ has been chosen to be *white*. For channel estimation, this is actually to be preferred, as a white $\{s(t)\}$ is optimal in terms of exciting the channel [80].

To be able to form a good estimate $\hat{\mathbf{h}}_L^{(n)}$ of a sparse channel we should know an appropriate value of n and which elements in $\hat{\mathbf{h}}_L^{(n)}$ should be set to zero. The choice of L is not very critical for the computational complexity of the method we will present and L should thus be chosen "large enough". The problem of finding the structure of the channel can be treated in many different ways, e.g., via hypothesis testing or by using IC. In this chapter we present a simple way of obtaining several sparse candidate channels which are then averaged and used for equalization.

The Generalized IC (GIC) can, for complex valued data, be written in the form:

$$\mathrm{GIC}_L^{(n)} = -2\ln p_n(\mathbf{y}|\hat{\mathbf{h}}_L^{(n)}) + \nu(2n). \tag{7.2}$$

The first term of the GIC shows how well the known sequence of symbols convolved with the channel estimate fits the received data. For a linear channel in complex Gaussian noise, we have $-2\ln p_n(\mathbf{y}|\hat{\mathbf{h}}_L^{(n)}) = 2N\ln\hat{\sigma}^2 + C$, where $\hat{\sigma}^2 = \frac{1}{N}\sum_{t=0}^{N-1}|y(t) - \mathbf{s}^T(t)\hat{\mathbf{h}}_L^{(n)}|^2$ and C is a constant independent of n [69]. For the two well-known and widely used AIC (Akaike IC) and BIC (Bayesian IC) we have $\nu = 2$ and $\nu = \ln 2N$, respectively (for the complex valued data case). The conventional way of using GIC for channel selection is as follows: For some pre-specified interval of n, select the single non-sparse channel

$$\hat{\mathbf{h}}_n^{(n+1)} = \arg\min_n \mathrm{GIC}_n^{(n+1)}. \tag{7.3}$$

In what follows, we will consider BIC. The key to using BIC for channel averaging is the observation that the *posterior likelihood* of a channel $\mathbf{h}_L^{(n)}$, i.e. the probability of $\mathbf{h}_L^{(n)}$ given the data vector $\mathbf{y}$, can be estimated directly using BIC in the following manner [69]:

$$\hat{P}(\mathbf{h}_L^{(n)}|\mathbf{y}) = \frac{e^{-\frac{1}{2}\mathrm{BIC}_L^{(n)}}}{\sum e^{-\frac{1}{2}\mathrm{BIC}}} \tag{7.4}$$

where the sum is over *all* channels considered. By using these posterior probabilities as weights we can form a weighted sum over all channels considered to obtain a channel average $\hat{\hat{\mathbf{h}}}_L^{(\bar{n})}$ (see below).

7.2 Sparse Channel Averaging

Model averaging for non-sparse channels has already be considered in [78]. Herein we present a computationally efficient way of obtaining sparse channel structures via the estimation of $L+1$ channels with only one non-zero tap.

The assumption that the known input symbols $\{s(t)\}$ are uncorrelated means that the contribution of each tap to the measured signal $\mathbf{y}$ is decoupled from that of the other taps. Thus, we can compare the contribution of each tap separately and rank them in order of importance (see [81]).

Let $\mathbf{h}_{L,l}^{(1)}$ denote the channel $\mathbf{h}_L^{(1)}$ which has h_l as its single non-zero tap. By estimating the entry h_l of $\mathbf{h}_{L,l}^{(1)}$ via Least Squares (LS) (or ML under the assumptions made)

$$\hat{h}_l = \frac{\sum_{t=0}^{N-1} s^*(t-l)y(t)}{\sum_{t=0}^{N-1} |s(t-l)|^2}, \quad l = 0, \ldots, L \tag{7.5}$$

($(\cdot)^*$ denotes the complex conjugate) we obtain $L+1$ one-tap ML channel estimates. Using these channel estimates we can compute the corresponding residual variances

$$\hat{\sigma}_l^2 = \frac{1}{N} \sum_{t=0}^{N-1} |y(t) - s(t-l)\hat{h}_l|^2, \quad l = 0, \ldots, L \tag{7.6}$$

and compute an estimated channel $\hat{\mathbf{h}}_L^{(n)}$ for each number of non-zero taps n up to a user selected limit $\bar{n} \leq L+1$, such that the positions of the non-zero taps of $\hat{\mathbf{h}}_L^{(n)}$ correspond to the n smallest values of $\{\hat{\sigma}_l^2\}_{l=0}^{L}$. This amounts to choosing the n taps most important for describing the received signal and is equivalent to selecting the taps corresponding to the n smallest values of $\text{GIC}_L^{(1)}$ for all $L+1$ one-tap channels, due to the common factor 2ν (as $n=1$). Once a channel structure for each $n = 1, \ldots, \bar{n}$ has been decided upon, we can find the ML/LS estimates $\{\hat{\mathbf{h}}_L^{(n)}\}_{n=1}^{\bar{n}}$ and average them using (7.2) and (7.4) in the following manner, such that the *averaged channel estimate* $\hat{\hat{\mathbf{h}}}_L^{(\bar{n})}$ is given by:

$$\hat{\hat{\mathbf{h}}}_L^{(\bar{n})} = \sum_{n=1}^{\bar{n}} \hat{P}(\mathbf{h}_L^{(n)}|\mathbf{y}) \cdot \hat{\mathbf{h}}_L^{(n)}. \tag{7.7}$$

This averaged estimate of the channel can then be used for channel equalization.

7.3 Numerical Examples

In Fig. 7.1 we show the BER corresponding to a Viterbi equalizer for randomly generated 150 sample long BPSK symbol sequences. An ensemble of

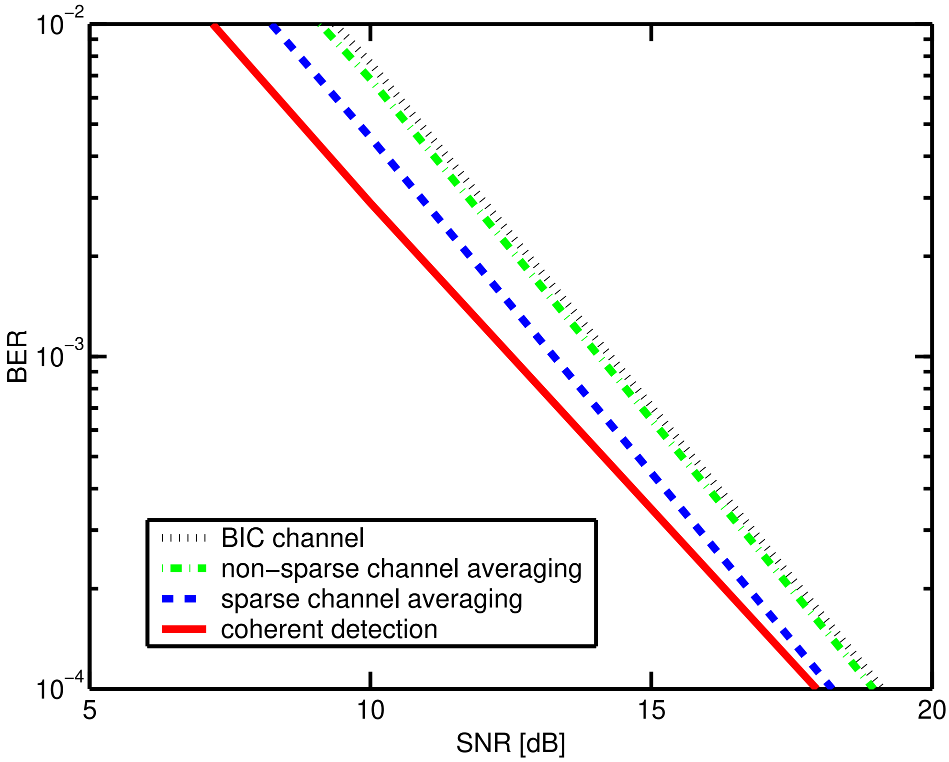


Figure 7.1: BER for the various channel estimation approaches.

Rayleigh fading channels with exponentially decaying delay profiles, a decay factor of $e^{-0.25}$ and a 50% probability for each of the $L + 1$ independent taps of being zero was used and circular Gaussian white noise was added to the channel outputs. Before equalization the channels were estimated using a known $N + L = 26$ sample BPSK GSM training sequence. L was set to 7 and $\bar{n}$ was set to $L + 1 = 8$. 10000 Monte Carlo simulations were run for several SNRs. The channels were estimated using the standard BIC, a non-sparse channel averaging approach presented in [78] and the sparse channel averaging approach presented here. Finally coherent detection was employed, where the true channel was used for equalization. We observe that an improvement of about 1 dB is obtained when using our sparse channel averaging approach instead of the standard BIC or the non-sparse channel averaging.

Bibliography

[1] T. Söderström and P. Stoica, *System Identification.* London, UK: Prentice Hall International, 1989.

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

[3] A. D. R. McQuarrie and C.-L. Tsai, *Regression and Time Series Model Selection.* Singapore: World Scientific Publishing, 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, 2002.

[6] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics.* Tokyo: KTK, 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.* New York: Springer-Verlag, 2nd ed., 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, England: Academic Press, 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.* Upper Saddle River, NJ: Prentice Hall, 2nd ed., 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* (Haykin, Litva, and Shepherd, eds.), pp. 99–151, Berlin: Springer-Verlag, 1993.

[18] P. Stoica and A. Nehorai, "Performance Study of Conditional and Unconditional Direction-of-Arrival Estimation," *IEEE Trans. ASSP*, vol. 38, pp. 1783–1795, October 1990.

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

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

[21] P. Djurić, "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: John Wiley and Sons, Inc., 1968.

[23] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, March 1951.

[24] H. Akaike, "A New Look at Statistical Model Identification," *IEEE Trans. Automatic Control*, vol. 19, pp. 716–723, 1974.

[25] H. Akaike, "On the Likelihood of a Time Series Model," *The Statistician*, vol. 27, pp. 217–235, 1978.

[26] R. L. Kashyap, "Inconsistency of the AIC Rule for Estimating the Order of AR Models," *IEEE Trans. Automatic Control*, vol. 25, pp. 996–998, 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," *Statistics and Probability Letters*, vol. 33, pp. 201–208, April 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, February 2003.

[31] P. M. T. Broersen, "Finite Sample Criteria for Autoregressive Order Selection," *IEEE Trans. Signal Processing*, vol. 48, pp. 3550–3558, December 2000.

[32] P. M. T. Broersen, "Automatic Spectral Analysis with Time Series Models," *IEEE Trans. on Instrumentation and Measurement*, vol. 51, pp. 211–216, April 2002.

[33] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, 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, pp. 99–104, 1982.

[35] J. Rissanen, "Modeling by Shortest Data Description," *Automatica*, vol. 14, pp. 465–471, 1978.

[36] J. Rissanen, "Estimation of Structure by Minimum Description Length," *Circuits, Syst., Sign. Process.*, vol. 1, pp. 395–406, 1982.

[37] T. W. Anderson, "The choice of the degree of a polynomial regression as a multiple decision problem," *Ann. of Math. Stat.*, vol. 33, pp. 255–265, 1962.

[38] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109–1113, 1989.

[39] M. Stone, "Cross-validatory choice and assessment of statistical predictions (with discussion)," *Journal of the Royal Statistical Society B*, vol. 36, pp. 111–147, 1974.

[40] J. Rissanen, "A predictive least-squares principle," *IMA J. Math. Contr. Inform.*, vol. 3, pp. 211–222, 1986.

[41] M. Wax, "Order selection for AR models by predictive least squares," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 581–588, 1988.

[42] H. Akaike, "Fitting Autoregressive Models for Prediction," *Ann. Inst. Statist. Math.*, vol. 21, pp. 243–247, 1969.

[43] H. Akaike, "An Objective use of Bayesian Models," *Ann. Inst. Statist. Math.*, vol. 29, pp. 9–20, 1977.

[44] F. Gustafsson and H. Hjalmarsson, "Twenty-one ML Estimators for Model Selection," *Automatica*, vol. 31, pp. 1377–1392, 1995.

[45] P. Stoica, "Multistep prediction of autoregressive signals," *Electronics Letters*, vol. 29, pp. 554–555, 1993.

[46] P. Stoica and Y. Selén, "Model-Order Selection — A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, pp. 36–47, July 2004.

[47] C. Gourieroux and A. Monfort, *Statistics and Econometric Models*, vol. I and II. Cambridge: Cambridge University Press, 1995.

[48] S. M. Kay, *Fundamentals of Statistical Signal Processing*, vol. 2: Detection Theory. Upper Saddle River, NJ: Prentice Hall, 1998.

[49] T. Söderström, "On model structure testing in system identification," *International Journal of Control*, vol. 26, no. 1, pp. 1–18, 1977.

[50] B. M. Pötscher, "Effects of Model Selection on Inference," *Econometric Theory*, vol. 7, pp. 163–185, 1991.

[51] H. Gu, "Estimating the Number of Signals and Signal Resolution," *IEEE Transactions on Signal Processing*, vol. 46, pp. 2267–2270, August 1998.

[52] W. Jiang and X. Liu, "Consistent model selection based on parameter estimates," *Journal of Statistical Planning and Inference*, 2004, to appear.

[53] D. S. Poskitt and A. R. Tremayne, "Determining a Portfolio of Linear Time Series Models," *Biometrika*, vol. 74, no. 1, pp. 125–137, 1987.

[54] L. Wasserman, "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology*, vol. 44, pp. 92–107, March 2000.

[55] J. A. Hoeting, D. Madigan, A. E. Raferty, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, vol. 14, pp. 382–417, November 1999.

[56] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, pp. 773–795, June 1995.

[57] A. C. Singer and M. Feder, "Universal Linear Prediction by Model Order Weighting," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2685–2699, October 1999.

[58] S. Sando, A. Mitra, and P. Stoica, "On the Cramér-Rao Bound for Model-Based Spectral Analysis," *IEEE Signal Processing Letters*, vol. 9, pp. 68–71, February 2002.

[59] J. G. Proakis, *Digital Communications.* New York: McGraw-Hill, Inc., 4th ed., 2000.

[60] H. Meyr, M. Moeneclaey, and S. A. Fechtel, *Digital Communication Receivers: Synchronization, Channel Estimation and Signal Processing.* New York: John Wiley & Sons, 1998.

[61] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, April 1967.

[62] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Transactions on Information Theory*, vol. 18, pp. 363–378, May 1972.

[63] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications.* Olympia, WA: Telecom Publishing, 1992.

[64] ETSI, "Digital cellular telecommunications system: Radio transmission and reception (GSM 05.05)." ETSI standardization document, 2002.

[65] H. L. Kazecki, S. H. Goode, D. W. Dennis, J. C. Baker, K. L. Baum, and B. D. Mueller, "Method for channel adaptive detecting/equalizing." U. S. Patent no. 5,195,106, March 16 1993.

[66] A. A. Abd El-Sallam and A. M. Zoubir, "A frequency domain based approach for low complexity rake receivers in cdma systems," in *Statistical Signal Processing, 2003 IEEE Workshop on*, (St. Louis, Missouri, USA), pp. 327–330, 28 September - 1 October 2003.

[67] D. T. Magill, "Optimal adaptive estimation of sampled stochastic processes," *IEEE Transactions on Automatic Control*, vol. 10, pp. 434–439, October 1965.

[68] W. S. Chaer and J. Bishop, R. H.and Ghosh, "A mixture-of-experts framework for adaptive Kalman filtering," *IEEE Transactions on Systems, Man and Cybernetics–Part B: Cybernetics*, vol. 27, pp. 452–464, June 1997.

[69] P. Stoica, Y. Selén, and J. Li, "Multi-model approach to model selection," *Digital Signal Processing*, 2004. in press.

[70] P. Stoica and Y. Selén, "Cross-Validation Rules for Order Estimation," *Digital Signal Processing*, vol. 14, no. 4, pp. 355–371, 2004.

[71] A. D. Lanterman, “Schwarz and Wallace and Rissanen: Intertwining themes in theories of model order estimation,” *International Statistical Review*, vol. 69, pp. 185–212, August 2001.

[72] L. Xu, A. Krzyzak, and C. Y. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, pp. 418–435, May-June 1992.

[73] R. A. Redner and H. F. Walker, “Mixture densities, maximum likelihood and the EM algorithm,” *SIAM Review*, vol. 26, pp. 195–239, April 1984.

[74] J. M. Wozencraft and I. M. Jacobs, *Principles of communications engineering*. Waveland Press, 1965.

[75] H. V. Poor, *An introduction to signal detection and estimation*. New York: Springer Verlag, 1988.

[76] E. G. Larsson, P. Stoica, and J. Li, “On maximum-likelihood detection and decoding for space-time coding systems,” *IEEE Transactions on Signal Processing*, vol. 50, pp. 937–944, April 2002.

[77] J. Hagenauer and P. Hoeher, “A Viterbi algorithm with soft-decision outputs and its applications,” in *Proc. of the IEEE Global Communications Conference*, vol. 3, pp. 1680–1686, 27-30 November 1989.

[78] E. G. Larsson, Y. Selén, and P. Stoica, “Adaptive equalization for frequency-selective channels of unknown length,” in *IEEE Global Telecommunications Conference (Globecom)*, (Dallas, Texas, USA), November 29 – December 3 2004. accepted for publication.

[79] E. G. Larsson, Y. Selén, and P. Stoica, “Adaptive equalization for frequency-selective channels of unknown length,” *IEEE Transactions on Vehicular Technology*, 2004. Submitted.

[80] O. Besson and P. Stoica, “Data-aided frequency offset estimation in frequency selective channels: training sequence selection,” in *ICASSP '02*, vol. 3, (Orlando, USA), pp. 2445–2448, 13-17 May 2002.

[81] J. Homer, I. Mareels, R. R. Bitmead, B. Wahlberg, and F. Gustafsson, “LMS Estimation via Structural Detection,” *IEEE Transactions on Signal Processing*, vol. 46, pp. 2651–2663, October 1998.

Recent licentiate theses from the Department of Information Technology

2003-006 Tomas Olsson: *Bootstrapping and Decentralizing Recommender Systems*

2003-007 Maria Karlsson: *Market Based Programming and Resource Allocation*

2003-008 Zoran Radovic: *Efficient Synchronization and Coherence for Nonuniform Communication Architectures*

2003-009 Martin Karlsson: *Cache Memory Design Trade-offs for Current and Emerging Workloads*

2003-010 Dan Wallin: *Exploiting Data Locality in Adaptive Architectures*

2003-011 Tobias Amnell: *Code Synthesis for Timed Automata*

2003-012 Olivier Amoignon: *Adjoint-Based Aerodynamic Shape Optimization*

2003-013 Stina Nylander: *The Ubiquitous Interactor - Mobile Services with Multiple User Interfaces*

2003-014 Kajsa Ljungberg: *Numerical Methods for Mapping of Multiple QTL*

2003-015 Erik Berg: *Methods for Run Time Analysis of Data Locality*

2004-001 Niclas Sandgren: *Parametric Methods for Frequency-Selective MR Spectroscopy*

2004-002 Markus Nordén: *Parallel PDE Solvers on cc-NUMA Systems*

2004-003 Yngve Selén: *Model Selection*

UPPSALA
UNIVERSITET

Department of Information Technology, Uppsala University, Sweden