Model Selection Based on Bayesian Predictive Densities and Multiple Data Records

Petar M. Djurić, *Member, IEEE,* and Steven M. Kay, *Fellow, IEEE*

Abstract— **Bayesian predictive densities are used to derive model selection rules. The resulting rules hold for sets of data records where each record is composed of an unknown number of deterministic signals common to all the records and a stationary white Gaussian noise. To determine the correct model, the set of data records is partitioned into two disjoint subsets. One of the subsets is used for *estimation* of the model parameters and the remaining for *validation* of the model. Two proposed estimators for linear nested models are examined in detail and some of their properties derived. Optimal strategies for partitioning the data records into estimation and validation subsets are discussed and analytical comparisons with the information criterion A of Akaike (AIC) and the minimum description length (MDL) of Schwarz and Rissanen are carried out. The performance of the estimators and their comparisons with the AIC and MDL selection rules are illustrated by numerical simulations. The results show that the Bayesian selection rules outperform the popular AIC and MDL criteria.**

I. Introduction

MODEL selection is an important problem in a variety of scientific and engineering disciplines. In signal processing it is equivalent to the detection of the number of signals in a multichannel time series [34]; the determination of a filter order in adaptive estimation [6]; pole retrievement of a system from its natural response [20]; speech, image and biomedical data compression [7], [22], [27]; segmentation of time series and digital images [29]. In classical statistics this problem is addressed by multiple hypotheses testing, which often cannot be handled easily since it requires the choice of a number of dependent significance levels. In addition, the multiple hypotheses testing may suffer from inconsistency and intransitivity [15]. Recently instead, the model selection problem has been pursued by using approaches founded on information theoretic [3], Bayesian [16], [28], and coding theoretic [25] reasoning.

In the paper we derive Bayesian model selection rules from multiple data records. The rules rest on the use of *predictive densities* according to the examined models *and* one portion of the observed data [1], [26]. The main idea is to partition the data into two subsets. One is used for determination of the density function of the model parameters and the other for validation of the hypothesized model. The partitioning may be repeated, and the results of the estimation-validation steps appropriately combined. The approach is related to the cross-validation principle [31] or the predictive sample reuse method [11]. The difference is that we exploit an additional assumption about the data, that is, their probabilistic structure.

First we briefly discuss why we resort to estimation-validation. Then we derive selection rules for linear nested models. These rules turn out to be similar in form as the two most popular criteria for model selection, the AIC and MDL. They are represented by a sum of data and penalty terms. The penalty term is a function of the total number of data records, M, and the number of estimation data records, L. We prove several propositions for conditions under which the selection rules become consistent. In addition we prove that the probability of overparametrization is minimized when we use only one data record for estimation and the rest for validation. On the other hand, the probability of underparametrization is minimized when $M - 1$ data records are used for estimation, and one for validation. Since these are conflicting requirements, we develop a strategy of how to partition the data records. We find upper bounds of the probabilities of over—and underparametrization which can serve to determine an optimal strategy for estimation-validation. The results suggest that the optimal L is a function of the signal-to-noise ratio (SNR). The lower the SNR, the larger the value of L for improved performance. In a very broad range of SNRs, the optimal L is equal to one. Moreover, we show that the probability of overparametrization is not a function of the SNR. In the paper we also prove that asymptotically the Bayesian rule becomes equivalent to the AIC when $L = M - 1$. If $L = 1$, our rules are similar to the MDL, but they are not equivalent. All the theoretical results are supported by computer simulations. The comparisons with the AIC and the MDL show that the Bayesian rules, when optimal L is chosen (usually $L = 1$), produce best results.

The paper is organized as follows. First the problem will be stated. Then the two-data-record case will be examined, emphasizing the characteristic steps in deriving the predictive densities and the selection rule therefrom. Two model order estimators of special interest will be defined: the symmetric and the sequential. In Section IV these two estimators will be derived when there are M data records, L of them used for obtaining a proper distribution for the model parameters, and the remaining $M - L$ data records for validating the model through the derived predictive density. The question of how to

Manuscript received June 24, 1992; revised September 10, 1993. The associate editor coordinating the review of this paper and approving it for publication was Prof. Douglas Williams. This work was supported by the University of Rhode Island under graduate student Fellowship and the National Science Foundation under Award No. MIP-9110628.

P. M. Djuric is with the Electrical Engineering Department, State University of New York at Stony Brook, NY 11794 USA.

S. M. Kay is with the Electrical Engineering Department, University of Rhode Island, Kingston, RI 02881 USA.

IEEE Log Number 9401290.

partition the data set into two subsets, L and $M - L$, and the question of consistency of the estimators will be discussed in Section V. Finally, relations to other model selection schemes and concluding remarks will be given in Sections VI and VII.

II. Problem Formulation

We suppose that a linear system $\mathcal{S}$ of unknown order m has generated M independent sequences according to

$$\mathbf{y}_j = \mathbf{H}_m\theta_m + \mathbf{e}_j \qquad j = 1, 2, \cdots, M, \tag{1}$$

where $\mathbf{H}_m$ is an $N \times m$ deterministic matrix, θ_m is an $m \times 1$ vector of unknown deterministic system parameters, and $\mathbf{e}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We shall assume that σ^2 is known. The data sequences have the same number of samples N. Further, we suppose that there is a set of hierarchical models $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_q$ described by

$$\mathbf{y}_j = \mathbf{H}_k\theta_k + \mathbf{e}_j \qquad k = 1, 2, \cdots, q; \qquad j = 1, 2, \cdots, M, \tag{2}$$

where $q \geq m$, and $\mathbf{H}_k$ for $k = m$ in (2) is identical to $\mathbf{H}_m$ in (1). In essence, one of the models in the set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_q\}$ is identical to the true system $\mathcal{S}$.

The problem is to estimate the order of the system.

III. The Two-Data-Record Paste

We suppose that there are two independent sequences $\mathbf{y}_1$ and $\mathbf{y}_2$ generated by the same system $\mathcal{S}$ given by (1). Using the two sequences, we shall derive several estimators based on predictive densities.

Predictive densities are defined as marginal densities of observed data according to a model. For example, if the data vector $\mathbf{y}_1$ has an assumed density given the model and its parameters, $f(\mathbf{y}_1|\theta_k, \mathcal{M}_k)$, and the prior density of the model parameters is $f(\theta_k|\mathcal{M}_k)$, then the predictive density according to the model $\mathcal{M}_k$ is obtained by

$$f(\mathbf{y}_1|\mathcal{M}_k) = \int_{\Theta_k} f(\mathbf{y}_1|\theta_k, \mathcal{M}_k) f(\theta_k|\mathcal{M}_k) d\theta_k \tag{3}$$

where Θ_k is a k-dimensional parameter space specified by the model. If $p(\mathcal{M}_k|\mathbf{y}_1)$ is the posterior probability that the model $\mathcal{M}_k$ is correct given the data sequence $\mathbf{y}_1$, Bayes' theorem yields

$$p(\mathcal{M}_k|\mathbf{y}_1) = \frac{p(\mathcal{M}_k) f(\mathbf{y}_1|\mathcal{M}_k)}{f(\mathbf{y}_1)} \tag{4}$$

where $p(\mathcal{M}_k)$ is the a priori probability of the model $\mathcal{M}_k$ and $f(\mathbf{y}_1)$ the marginal density of the data.

If our criterion for model selection (or model order estimation) is the posterior probability of the model $\mathcal{M}_k$ after observing the data $\mathbf{y}_1$, $p(\mathcal{M}_k|\mathbf{y}_1)$, the best model maximizes (4). This criterion minimizes the probability of selection error $\mathcal{P}_\epsilon$ when the loss function is chosen such that it equals one for incorrect and zero for correct model selection [15]. Since $f(\mathbf{y}_1)$ is common for all the models, we are basically looking for the maximum of $p(\mathcal{M}_k) f(\mathbf{y}_1|\mathcal{M}_k)$. If we further assume that the a priori probabilities of the models are equal, then we choose the model which maximizes (3).

A major difficulty in using (3) is the quantification of the prior density $f(\theta_k|\mathcal{M}_k)$. If we decide to choose a noninformative prior,[1] then the posterior probabilities of the models may lead to arbitrary selections [4]. In the linear case, for example, the noninformative prior is an arbitrary constant which does not disappear while evaluating (3), necessarily leading to an arbitrary selection criterion [10]. On the other hand, if we use proper priors — for instance, conjugate priors[2]—then we have additional problems with assigning values to the distribution parameters, and these values for short data records may affect the selection [10].

We prefer to use the noninformative priors since they introduce minimal information. To alleviate the problem of arbitrariness we shall

1) use a noninformative prior to obtain a posterior density of θ_k using the first sequence $\mathbf{y}_1$, $f(\theta_k|\mathbf{y}_1, \mathcal{M}_k)$, which is proper; and
2) use $f(\theta_k|\mathbf{y}_1, \mathcal{M}_k)$ as a prior in (3) to obtain the predictive density of $\mathbf{y}_2$ according to the model $\mathcal{M}_k$ and $\mathbf{y}_1$.

Thus, the first sequence is used for determining a proper prior for the model parameters, which allows us to obtain the predictive density of the data $\mathbf{y}_2$ avoiding the arbitrary constants from the noninformative priors. Therefore, this is now a predictive density not only according to the model, but also to one portion of the observed data. Since the second data record $\mathbf{y}_2$ is also known, it can be substituted in the predictive density expression, yielding a measure of prediction accuracy of the examined model—i.e., validating it. For further convenience, the data records used for obtaining proper priors will be called estimation, and the remaining ones, validation data records.

Technically, the procedure is implemented as follows. We may write

$$f(\mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k) = \int_{\Theta_k} f(\mathbf{y}_2|\theta_k, \mathbf{y}_1, \mathcal{M}_k) f(\theta_k|\mathbf{y}_1, \mathcal{M}_k) d\theta_k \tag{5}$$

where $f(\mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k)$ is the predictive density of $\mathbf{y}_2$ according to the data $\mathbf{y}_1$ and the model $\mathcal{M}_k$. The predictive density as defined in (5) will be determined for every model in the set $\mathcal{M}$. We choose the model that maximizes (5).

For the linear models in (2) the noninformative priors are defined by [5]

$$f(\theta_k|\mathcal{M}_k) \propto \text{const.}$$

It is readily shown that the posterior distribution becomes [5]

$$f(\theta_k|\mathbf{y}_1, \mathcal{M}_k) = \frac{\left|\mathbf{H}_k^T\mathbf{H}_k\right|^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{k}{2}}}$$

$$\cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\theta_k - \hat{\theta}_k^{(1)}\right)^T \mathbf{H}_k^T\mathbf{H}_k\left(\theta_k - \hat{\theta}_k^{(1)}\right)\right\} \tag{6}$$

[1] Noninformative priors are defined by $f(\theta_k|\mathcal{M}_k) \propto |I(\theta_k)|^{\frac{1}{2}}$ - i.e., they are proportional to the determinant of Fisher's information matrix.

[2] Parametric priors which yield posterior densities that belong to the same family of prior densities are called conjugate priors.

where

$$\hat{\theta}_k^{(1)} = \left(\mathbf{H}_k^T\mathbf{H}_k\right)^{-1}\mathbf{H}_k^T\mathbf{y}_1.$$

Note that

$$f(\mathbf{y}_2|\theta_k, \mathbf{y}_1, \mathcal{M}_k) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_2 - \mathbf{H}_k\theta_k)^T(\mathbf{y}_2 - \mathbf{H}_k\theta_k)\right\}. \quad (7)$$

When (6) and (7) are substituted into (5), the nuisance parameters θ_k integrated out, and the irrelevant terms (which are identical for every k) dropped out, we obtain

$$f(\mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k) \propto \frac{1}{2^{\frac{k}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{1}{2}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 - \frac{1}{2}\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 - \mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_1\right)\right\} \quad (8)$$

where

$$\mathbf{P}_k = \mathbf{H}_k\left(\mathbf{H}_k^T\mathbf{H}_k\right)^{-1}\mathbf{H}_k^T \quad (9)$$

is a projection operator. From (8) we propose the following estimator for m

$$\hat{m}_{2|1} = \arg\left\{\min_k\left(-2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_1 - \mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 + \mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 + 2k\sigma^2\ln 2\right)\right\} \quad (10)$$

which is equivalent to selecting the model that maximizes (8). The index $2|1$ denotes that the sequence $\mathbf{y}_2$ was used for validation and $\mathbf{y}_1$ for estimation.

Clearly, we might have used the sequences in the opposite order and obtained a similar estimator

$$\hat{m}_{1|2} = \arg\min_k\left\{(-2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_1 - \mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 + \mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 + 2k\sigma^2\ln 2)\right\}. \quad (11)$$

Now we have two estimators which may yield different estimates using the same pair of sequences. Which estimator should be used then, or how should they be combined to obtain a unique estimate $\hat{m}$? (10) seems to be a natural choice if the data records are ordered in time. Otherwise, the following two alternatives seem to be reasonable choices

$$\hat{m}_{\min} = \min\left(\hat{m}_{1|2}, \hat{m}_{2|1}\right) \quad (12)$$

and

$$\hat{m}_{\text{sym}} = \arg\left\{\min_k(m_{1|2} + m_{2|1})\right\} = \arg\left\{\min_k\left(-2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_1 + 2k\sigma^2\ln 2\right)\right\}. \quad (13)$$

The estimator (12) is motivated by the principle of parsimony. We want to choose the simplest model allowed by the data, so we choose the simpler model from the two already selected by (10) and (11). The index "sym" in (13) stands for symmetrical because the form of the estimator is symmetrical with respect to $\mathbf{y}_1$ and $\mathbf{y}_2$ (each sequence is used as an estimation sequence and the remaining for validation). It is motivated by the desire: 1) to impose equal role to each of the sequences available for processing and 2) to decrease the variance of our estimates (note that (13) is based on the geometrical mean of all the predictive probability density functions that can be constructed from the data records.)

It is interesting that these estimators have similar formats like other model selection criteria (see Section VI). They have two terms, a data term and a penalizing term. For example, in (13) the data term represents the crosscorrelation of the observed vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ in the signal subspace represented by the projection operator $\mathbf{P}_k$. The second term is the penalizing factor which monotonically increases when the order of the model and the observation noise variance increase. This will entail the choice of simpler models whenever the crosscorrelations of $\mathbf{y}_1$ and $\mathbf{y}_2$ in the signal subspaces defined by the higher and lower dimension models are approximately equal.

Another interesting characteristic of the estimators (10)–(13) is that their penalty term is not a function of the data record length N. This will certainly imply that when $N \to \infty$, these estimators will not be consistent, which means that the probability of incorrectly selecting higher order models will remain finite.

The most stringent of the four estimators (10)–(13) in penalizing overparametrization is $\hat{m}_{\min}$. This can be shown analytically (omitted here) and by simulations. We have tested the estimators on 1000 trials using the following data model

$$y_t = \sum_{i=1}^{8} A_i\cos(2\pi f_i t + \phi_i) + e_t, \quad t = 1, 2, \cdots, N \quad (14)$$

where $A_1 = 1.89$, $A_2 = 1.41$, $A_3 = 1.06$, $A_k = 0$ for $k = 4, 5, 6, 7, 8$, and $\sigma^2 = 1$ (the correct model was 3). Note that now $\theta_k^T = [A_1 A_2 \cdots A_k]$. The rest of the parameters are given in Table I. The data record lengths were N= 20 and 50. The results are presented in Table II. The performance of the symmetric estimator has not improved with the increase of the data record lengths. The best performance (largest number of correct selections) achieved $\hat{m}_{\min}$. Even so, we conjecture that for the most part it may be overly conservative, notably for short data records or low SNRs. This was verified by many simulations. The results shown in Table III illustrate how the estimates obtained by $\hat{m}_{\min}$ deteriorated when the SNR was decreased compared to the case given in Table II. The same model was used as in (14) with identical parameters as before, except that $A_3 = 0.795$. In 25 cases the estimate was even $\hat{m}_{\min} = 1$, thus missing the two sinusoids at frequencies $f_2 = 0.15$ and $f_3 = 0.28$. When there are M sequences, this will be much more emphasized. Although parsimony is desirable, consistent underparametrization in many problems is much less acceptable than overparametrization. Therefore, this estimator will not be considered in the sequel.[3]

As already mentioned, we shall be interested primarily in deriving model selection criteria when there are more than two sequences. Two estimators will be analyzed. The first

[3]It should be noted, however, that the tendency of underparametrization can be reduced by imposing higher costs for underparametrization than for overparametrization.

TABLE I
PARAMETERS OF THE DATA MODEL (14); THE SAMPLING FREQUENCY IS 1

f_1	ϕ_1	f_2	ϕ_2	f_3	ϕ_3	f_4	ϕ_4
0.32	1.0	0.15	1.7	0.28	-0.4	0.10	2.1
f_5	ϕ_5	f_6	ϕ_6	f_7	ϕ_7	f_8	ϕ_8
0.44	-1.1	0.05	-2.4	0.21	0.9	0.38	1.3

TABLE II
COMPARISON OF ESTIMATORS WHEN THE DATA RECORDS HAD 20 AND 50 SAMPLES; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE SNR FOR THE THIRD SINUSOID IS –2.5 dB; THE CORRECT MODEL ORDER IS 3

	$M=2$, $N=20$				$M=2$, $N=50$			
k	$\hat{m}_{\text{sym}}$	$\hat{m}_{2\|1}$	$\hat{m}_{1\|2}$	$\hat{m}_{\min}$	$\hat{m}_{\text{sym}}$	$\hat{m}_{2\|1}$	$\hat{m}_{1\|2}$	$\hat{m}_{\min}$
1	0	1	0	0	0	0	0	0
2	0	54	36	91	0	3	6	9
3	736	610	594	769	731	635	648	837
4	116	109	131	76	125	138	140	83
5	32	53	50	12	70	84	87	44
6	59	83	96	35	36	57	44	15
7	36	50	41	12	25	37	39	7
8	21	40	52	5	13	46	36	5

TABLE III
COMPARISON OF ESTIMATORS WHEN THE DATA RECORDS HAD 20 SAMPLES; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE SNR FOR THE THIRD SINUSOID IS –5 dB; THE CORRECT MODEL ORDER IS 3

	$M=2$, $N=20$			
k	$\hat{m}_{\text{sym}}$	$\hat{m}_{2\|1}$	$\hat{m}_{1\|2}$	$\hat{m}_{\min}$
1	0	13	12	25
2	23	108	105	209
3	732	557	546	645
4	104	115	135	72
5	68	87	76	34
6	29	52	52	7
7	27	36	39	6
8	17	32	35	2

is the symmetric estimator, $\hat{m}_{\text{sym}}(M, L)$, where M denotes the total number of available data records, and L the number of estimation data records. (For instance, when $M = 2$ and $L = 1$, $\hat{m}_{\text{sym}}(2, 1)$ is given by (13).) The other estimator will be the sequential estimator, $\hat{m}_{\text{seq}}(M, L)$ (When $M = 2$ and $L = 1$, $\hat{m}_{\text{seq}}(2, 1)$ is given by (10).)

IV. THE GENERAL-DATA-RECORD CASE

When there are more then two data records, we have different possibilities in partitioning the data set into estimation and validation subsets. This partitioning is very important and deserves special attention. It will be investigated in Section V.

Before addressing the general case of M data records, we consider the problem when there are only three sequences $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$. To find $f(\mathbf{y}_3|\mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k)$, we write as before

$$f(\mathbf{y}_3|\mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k) = \int_{\Theta_k} f(\mathbf{y}_3|\theta_k, \mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k) \cdot f(\theta_k|\mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k) d\theta_k$$

where

$$f(\theta_k|\mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k) = \frac{\left|2\mathbf{H}_k^T\mathbf{H}_k\right|^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{k}{2}}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\theta_k - \hat{\theta}_k^{(1,2)}\right)^T 2\mathbf{H}_k^T\mathbf{H}_k\left(\theta_k - \hat{\theta}_k^{(1,2)}\right)\right\}$$

and

$$\hat{\theta}_k^{(1,2)} = \frac{1}{2}\left(\mathbf{H}_k^T\mathbf{H}_k\right)^{-1}\mathbf{H}_k^T(\mathbf{y}_1 + \mathbf{y}_2).$$

$f(\mathbf{y}_3|\theta_k, \mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k)$ has a form similar to $f(\mathbf{y}_1|\theta_k, \mathbf{y}_2, \mathcal{M}_k)$ in (7). After some algebra, we find that

$$f(\mathbf{y}_3|\mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k) \propto \left(\frac{2}{3}\right)^{\frac{k}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{1}{6}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 + \frac{1}{6}\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 - \frac{1}{3}\mathbf{y}_3^T\mathbf{P}_k\mathbf{y}_3 + \frac{1}{3}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_2 - \frac{2}{3}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_3 - \frac{2}{3}\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_3\right)\right\}. \quad (15)$$

Thus, from (15) we deduce that the sequential estimator is

$$\hat{m}_{\text{seq}}(3, 2) = \arg\left\{\min_k \left(-2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_3 - 2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_3 + \mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_2 + \frac{1}{2}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 + \frac{1}{2}\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 - \mathbf{y}_3^T\mathbf{P}_k\mathbf{y}_3 + 3k\sigma^2\ln\frac{3}{2}\right)\right\}. \quad (16)$$

After symmetrization, i.e., taking each possible combination of data records for estimation and validation, the following estimator is obtained

$$\hat{m}_{\text{sym}}(3, 2) = \arg\left\{\min_k \left(-2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_2 - 2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_3 - 2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_3 + 6k\sigma^2\ln\frac{3}{2}\right)\right\}. \quad (17)$$

Next, we derive the expression for $f(\mathbf{y}_3, \mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k)$. Note that

$$f(\mathbf{y}_3, \mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k) = f(\mathbf{y}_3|\mathbf{y}_2, \mathbf{y}_1, \mathcal{M}_k) f(\mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k).$$

Since the two terms on the right-hand side have already been evaluated, we easily find that

$$f(\mathbf{y}_3, \mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k) \propto \left(\frac{1}{3}\right)^{\frac{k}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{2}{3}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 - \frac{1}{3}\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 - \frac{1}{3}\mathbf{y}_3^T\mathbf{P}_k\mathbf{y}_3 - \frac{2}{3}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_2 - \frac{2}{3}\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_3 - \frac{2}{3}\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_3\right)\right\}.$$

Therefore

$$\hat{m}_{\text{seq}}(3, 1) = \arg\left\{\min_k \left(-2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_2 - 2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_3 - 2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_3 + 2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_1 - \mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_2 - \mathbf{y}_3^T\mathbf{P}_k\mathbf{y}_3 + 3k\sigma^2\ln 3\right)\right\}. \quad (18)$$

and

$$\hat{m}_{\text{sym}}((3, 1)) = \arg\left\{\min_k \left(-2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_2 - 2\mathbf{y}_1^T\mathbf{P}_k\mathbf{y}_3 - 2\mathbf{y}_2^T\mathbf{P}_k\mathbf{y}_3 + 3k\sigma^2\ln 3\right)\right\}. \quad (19)$$

In a comparison of (16) and (18) it is not easy to see the difference between the two estimators, while in comparing (17) and (19), we notice that the estimators differ only in the penalizing term. The data terms in (17) and (19) represent all the combinations of crosscorrelation between the data records in the signal subspace, while the autocorrelations are excluded as before.

For a given $M > 3$ we may construct $M - 1$ different symmetric and sequential estimators, depending on how many data records will be used for estimation and validation.[4] In general, the derivation closely follows the lines of reasoning for the two- and three-data-record cases. It is given in Appendix A. It is shown that

$$\hat{m}_{\text{sym}}(M, L) = \arg\left\{\min_k \left(-\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k (\mathbf{y}^{(M)} - \mathbf{y}_j) + k\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}\right)\right\}$$

where

$$\mathbf{y}^{(M)} = \sum_{i=1}^{M} \mathbf{y}_j.$$

Another form of this estimator that will be convenient for later use is

$$\hat{m}_{\text{sym}}(M, L) = \arg\left\{\min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k \mathbf{y}_j + k\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}\right)\right\}. \quad (20)$$

An interesting and obvious property of $\hat{m}_{\text{sym}}(M, L)$ is the following:

Property 1: The data term is not a function of L.

Another property that is related to the penalization term,

$$P(M, L) = k\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}$$

is

Property 2: The penalization function $P(M, L)$ for fixed M decreases monotonically with L.

This property can be proved readily using the inequality

$$\ln z < z - 1, \quad z > 1.$$

Thus, the most stringent penalization for overparametrization will be obtained if $L = 1$, and the weakest for $L = M - 1$. This implies that when fewer sequences are used for determining the initial proper prior distribution of the model parameters, higher model order estimates will be less likely, and vice versa.

[4] Actually the total number of different sequential estimators is $\sum_{L=1}^{M-1} \frac{M!}{(M-L)!L!}$. But if the data records are ordered (in time, for example), and we suppose that they will be processed in that order, then there will be $M - 1$ different sequential estimators.

Now we turn the attention to the sequential estimator. In Appendix A it is shown that its form is

$$\hat{m}_{\text{seq}}(M, L) = \arg\left\{\min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \frac{M}{L}\mathbf{y}^{(L)^T} \mathbf{P}_k \mathbf{y}^{(L)} + kM\sigma^2 \ln\frac{M}{L}\right)\right\}. \quad (21)$$

The corresponding properties of this estimator are the following:

Property 1: The data term is a function of L.

Property 2: The penalty term decreases monotonically as L increases.

When compared to the penalty in (20), we deduce that the penalty term in (21) is less stringent for $L > 1$, while for $L = 1$, they are identical. This subject will be discussed further in the next two sections.

V. Strategy for Estimation-Validation

There are two important questions to be addressed concerning estimators (20) and (21). They are

1) Are these estimators consistent when the number of data records M increases?
2) For fixed M, how do we choose L?

The answer to the first question is given by the following two propositions:

Proposition 1: If the assumptions in Section II are true, L is fixed, and $M \to \infty$, then the estimator (20) is consistent.

Proof: See Appendix B.

Proposition 2: If the assumptions in Section II are true, $M - L$ is fixed, and $M \to \infty$, then the estimator (26) is not consistent.

Proof: The proof closely follows the proof of Proposition 1.

It can be shown that under the conditions stated in Proposition 2 the probability for underparametrization $\mathcal{P}(\hat{m} < m) \to 0$ as $M \to \infty$, while the probability for overparametrization $\mathcal{P}(\hat{m} > m)$ remains finite.

Analogous propositions can be stated for the sequential estimator.

In Table IV some results are shown which agree with Propositions 1 and 2. The symmetric estimators $\hat{m}_{\text{sym}}(M, 1)$ and $\hat{m}_{\text{sym}}(M, M - 1)$ were examined on data generated as in (14). M was varied from 10 to 80. The performance of $\hat{m}_{\text{sym}}(M, 1)$ improved monotonically as M increased, while that of $\hat{m}_{\text{sym}}(M, M - 1)$ did not.

Now we address the strategy of estimation-validation, i.e., the issue of selection of number of data records for estimation. Consider first the symmetric estimators. Let

$$J_k(M, L) = -\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k \mathbf{y}_j + k\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}. \quad (22)$$

Then (20) can be rewritten as

$$\hat{m}_{\text{sym}}(M, L) = \arg\left\{\min_k (J_k(M, L))\right\}, \qquad 1 \le k \le q.$$

We state the following proposition:

TABLE IV
COMPARISON OF PERFORMANCE OF SYMMETRIC ESTIMATORS; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE CORRECT MODEL ORDER IS 3

	$\hat{m}_{sym}(M,1)$				$\hat{m}_{sym}(M,M-1)$			
k	$M=10$	$M=20$	$M=40$	$M=80$	$M=10$	$M=20$	$M=40$	$M=80$
1	0	0	0	0	0	0	0	0
2	266	80	7	0	118	16	0	0
3	668	869	968	986	627	711	752	716
4	42	36	20	14	121	130	95	126
5	18	12	4	0	68	58	61	67
6	2	3	1	0	24	44	33	34
7	4	0	0	0	19	20	33	35
8	0	0	0	0	23	21	26	22

Proposition 3: The probability of overparametrization, $\mathcal{P}_o(L)$, is minimized when $L = 1$, and the probability of underparametrization, $\mathcal{P}_u(L)$, is minimized when $L = M-1$.

Proof: For the first part of the proof it will be enough to show that

$$\mathcal{P}_o(L) \leq \mathcal{P}_o(L+1).$$

Define the events $\mathcal{E}_{kl}(L)$ and $\mathcal{E}_k(L)$ according to

$$\mathcal{E}_{kl}(L): \quad J_k(M,L) < J_l(M,L), \qquad k > m, l \leq m$$

and

$$\begin{aligned}\mathcal{E}_k(L) &= \mathcal{E}_{k1}(L) \cap \mathcal{E}_{k2}(L) \cdots \cap \mathcal{E}_{km}(l) \\ &= \cap_{l=1}^{m} \mathcal{E}_{kl}(L), \qquad k > m. \end{aligned} \tag{23}$$

The event of overparametrization $\mathcal{E}_o(L)$ is then the union of events $\mathcal{E}_k(L)$, for $k > m$, i.e.

$$\begin{aligned}\mathcal{E}_o(L) &= \mathcal{E}_{m+1}(L) \cup \mathcal{E}_{m+2}(L) \cdots \cup \mathcal{E}_q(L) \\ &= \cup_{k=m+1}^{q} \mathcal{E}_k(L). \end{aligned} \tag{24}$$

First we need to show that

$$\mathcal{E}_{kl}(L) \subset \mathcal{E}_{kl}(L+1), \quad k > m, l \leq m.$$

From (23) we have

$$\mathcal{E}_{kl}(L): \quad -\mathbf{y}^{(M)^T}\mathbf{P}_k\mathbf{y}^{(M)} + \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_k\mathbf{y}_j$$

$$+ k\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L} < -\mathbf{y}^{(M)^T}\mathbf{P}_l\mathbf{y}^{(M)} + \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_l\mathbf{y}_j$$

$$+ l\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}, \qquad k > m, l \leq m$$

or

$$\mathcal{E}_{kl}(L): \quad -\mathbf{y}^{(M)^T}\mathbf{P}_{kl}^{\perp}\mathbf{y}^{(M)} + \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_{kl}^{\perp}\mathbf{y}_j$$

$$< (l-k)\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}, \qquad k > m, l \leq m$$

where

$$\mathbf{P}_{kl}^{\perp} = \mathbf{P}_k - \mathbf{P}_l.$$

Note that the left hand side of the inequality is not a function of L, while the right hand side is a monotonically increasing function of L (see Property 2 of the symmetric estimator). It clearly implies

$$\mathcal{E}_{kl}(L) \subseteq \mathcal{E}_{kl}(L+1), \quad k > m, l \leq m.$$

Secondly, this relationship entails

$$\begin{aligned}\mathcal{E}_k(L) &= \cap_{l=1}^{m} \mathcal{E}_{kl}(L) \subseteq \mathcal{E}_k(L+1) \\ &= \cap_{l=1}^{m} \mathcal{E}_{kl}(L+1) \quad k > m. \end{aligned} \tag{25}$$

Consequently, for the event defined by (24), and using (25), we can write

$$\mathcal{E}_o(L) = \bigcup_{k=m+1}^{q} \mathcal{E}_k(L) \subseteq \mathcal{E}_o(L+1) = \cup_{k=m+1}^{q} \mathcal{E}_k(L+1)$$

or

$$P_o(L) = P(\mathcal{E}_o(L)) \leq P(\mathcal{E}_o(L+1)) = P_o(L+1).$$

The last expression implies the claim of the first part of the proposition. The proof for the second part follows the same lines as for the first part. ◇

Clearly, we have two conflicting conditions on L for minimization of the probabilities of under- and overparametrization when the minimization is carried out independently. By contrast, we would like to choose L when the two types of errors are considered simultaneously. Therefore, we want to find the probability of error, $\mathcal{P}_\epsilon$, defined by

$$\mathcal{P}_\epsilon = \mathcal{P}_o + \mathcal{P}_u, \tag{26}$$

and examine for which L it is minimized. It turns out that the determination of the exact form of $\mathcal{P}_\epsilon$ is very difficult because we have to work with correlated random variables defined by $\mathbf{y}_i^T\mathbf{P}_k\mathbf{y}_j$, $i \neq j$, whose density functions are unknown. For this reason we settle for less and try to determine the upper bound of the probability of error. Without loss of generality let $m > 1$. To simplify the mathematics, we have three additional assumptions

$$\mathcal{P}(\hat{m} = k > m) = \mathcal{P}(J_{m+1}(M,L) < J_m(M,L)) \tag{27}$$

$$\mathcal{P}(\hat{m} = k < m) = \mathcal{P}(J_{m-1}(M,L) < J_m(M,L)) \tag{28}$$

and

$$\mathcal{P}((J_k(M,L) < J_m(M,L)) \cap (J_l(M,L) < J_m(M,L))) = 0 \tag{29}$$

for $k > m$, and $l < m$. All these assumptions will be true if $J_k(M,L)$ as a function of k has only one minimum when k varies from 1 to q. The performance analysis of the information theoretic criteria was carried out with similar approximations in [17], [33], and [35].

The following random variables are now of particular interest

$$r_o = -\mathbf{y}^{(M)^T}\mathbf{P}_{m+1}\mathbf{y}^{(M)} + \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_{m+1}\mathbf{y}_j$$

$$+ \mathbf{y}^{(M)^T}\mathbf{P}_m\mathbf{y}^{(M)} - \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_m\mathbf{y}_j$$

and

$$r_u = -\mathbf{y}^{(M)^T}\mathbf{P}_{m-1}\mathbf{y}^{(M)} + \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_{m-1}\mathbf{y}_j$$

$$+ \mathbf{y}^{(M)^T}\mathbf{P}_m\mathbf{y}^{(M)} - \sum_{j=1}^{M}\mathbf{y}_j^T\mathbf{P}_m\mathbf{y}_j. \tag{30}$$

Now according to assumptions (27) and (29)

$$\mathcal{P}_o = \mathcal{P}\left(r_o < -\frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}\right)$$

and

$$\mathcal{P}_u = \mathcal{P}\left(r_u < \frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}\right).$$

In order to make any progress, we need to know something about the statistics of r_o and r_u. It turns out that we can determine their moments. To find them, we represent r_0 and r_u as quadratic forms, viz.

$$\begin{aligned} r_o &= -\check{\mathbf{y}}^T\check{\mathbf{P}}_{m+1}\check{\mathbf{y}} + \check{\mathbf{y}}^T\check{\mathbf{P}}_m\check{\mathbf{y}} \\ &= -\check{\mathbf{y}}^T\check{\mathbf{P}}_o\check{\mathbf{y}} \end{aligned}$$

and

$$\begin{aligned} r_u &= -\check{\mathbf{y}}^T\check{\mathbf{P}}_m\check{\mathbf{y}} + \check{\mathbf{y}}^T\check{\mathbf{P}}_{m-1}\check{\mathbf{y}} \\ &= \check{\mathbf{y}}^T\check{\mathbf{P}}_u\check{\mathbf{y}} \end{aligned}$$

where

$$\check{\mathbf{y}}^T = [\mathbf{y}_1^T\mathbf{y}_2^T\cdots\mathbf{y}_M^T]$$

$$\check{\mathbf{P}}_k = \begin{pmatrix} \mathbf{0} & \mathbf{P}_k & \cdots & \mathbf{P}_k \\ \mathbf{P}_k & \mathbf{0} & \cdots & \mathbf{P}_k \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}_k & \mathbf{P}_k & \cdots & \mathbf{0} \end{pmatrix}, \quad k = m-1, m, m+1$$

$$\check{\mathbf{P}}_o = \begin{pmatrix} \mathbf{0} & \mathbf{P}_o & \cdots & \mathbf{P}_o \\ \mathbf{P}_o & \mathbf{0} & \cdots & \mathbf{P}_o \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}_o & \mathbf{P}_o & \cdots & \mathbf{0} \end{pmatrix}$$

$$\check{\mathbf{P}}_u = \begin{pmatrix} \mathbf{0} & \mathbf{P}_u & \cdots & \mathbf{P}_u \\ \mathbf{P}_u & \mathbf{0} & \cdots & \mathbf{P}_u \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}_u & \mathbf{P}_u & \cdots & \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{P}_o = \mathbf{P}_{m+1} - \mathbf{P}_m.$$

$$\mathbf{P}_u = \mathbf{P}_m - \mathbf{P}_{m-1}.$$

Next we state the following proposition:

Proposition 4: The expected values of r_o and r_u are given by

$$E(r_o) = 0$$

and

$$E(r_u) = M(M-1)\theta_m^T\mathbf{H}_m^T\mathbf{P}_u\mathbf{H}_m\theta_m.$$

Proof: The results follow from the identity

$$E(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{A}\mathbf{C}) + \mu^T\mathbf{A}\mu$$

where $\mu = E(\mathbf{x})$, and $\mathbf{C}$ is the covariance matrix of $\mathbf{x}$. ◇

In addition, we shall use the following result:

Proposition 5: The l-th cumulant of r_o and r_u is given by

$$K_l = 2^{l-1}(l-1)!\,\sigma^{2l}\,\mathrm{tr}(\check{\mathbf{P}})^l$$

where $\check{\mathbf{P}}$ is equal to $\check{\mathbf{P}}_u$ and $\check{\mathbf{P}}_o$, respectively.

Proof: The result follows from a theorem which states that the lth cumulant of $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is given by [18], [30]

$$K_l = 2^{l-1}(l-1)!(\mathrm{tr}(\mathbf{A}\mathbf{C})^l + l\mu^T\mathbf{A}(\mathbf{C}\mathbf{A})^{l-1}\mu).$$

◇

Now we are ready to state the following two propositions:

Proposition 6: The Cantelli upper bound of the probability of overparametrization, $\mathcal{P}_o$, is given by

$$\mathcal{P}_0 \le \frac{1}{1 + \frac{M(M-1)\ln^2(\frac{M}{L})}{2(M-L)^2}}. \tag{31}$$

Proof: The bound is obtained by using Cantelli inequality [24], which for the random variable X states that

$$\mathcal{P}(X - \mu \le -\lambda) \le \frac{\sigma^2}{\sigma^2 + \lambda^2}, \quad \lambda > 0 \tag{32}$$

where μ and σ^2 are the mean value and the variance of X. In our case from Proposition 4 we have $E(r_o) = 0$ and from Proposition 5

$$K_2 = 2\sigma^4 M(M-1).$$

This implies that $\sigma_{r_o}^2 = 2\sigma^4 M(M-1)$. Since $\lambda = \frac{M(M-1)}{M-L}\sigma^2 \ln\frac{M}{L}$, we easily determine (31) from (32) using the obtained mean value and variance of r_o. ◇

Proposition 7: The Cantelli upper bound of the probability of underparametrization, $\mathcal{P}_u$, is given by

$$\mathcal{P}_u \le \frac{1}{1 + \frac{M^2(\eta - \frac{\ln\frac{M}{L}}{M-L})^2}{4(\eta + \frac{M-1}{2M})^2}}, \qquad \eta > \frac{\ln\frac{M}{L}}{M-L}, \tag{33}$$

where η is a "partial" SNR defined by

$$\eta = \frac{s^2}{\sigma^2},$$

where

$$s^2 = \theta_m^T\mathbf{H}_m^T\mathbf{P}_u\mathbf{H}_m\theta_m$$

and

$$\mathbf{P}_u = \mathbf{P}_m - \mathbf{P}_{m-1}.$$

◇

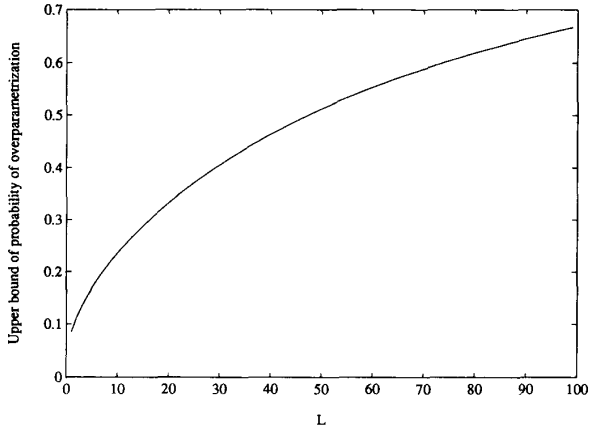

Fig. 1 Cantelli upper bound of probability of overparametrization as a function of L when $M = 100$.

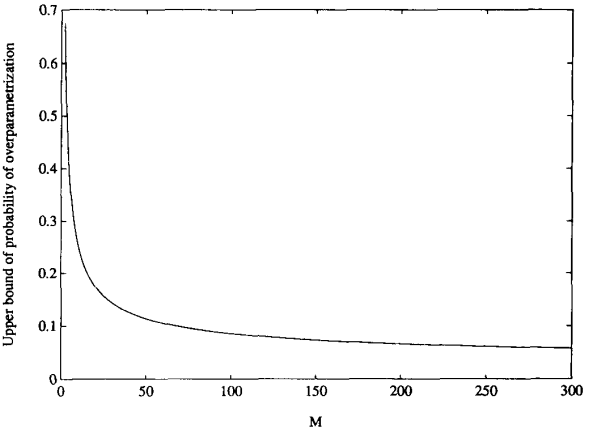

Fig. 2 Cantelli upper bound of probability of overparametrization as a function of M when $L = 1$.

Proof: The proof is similar to the previous one. We want to determine a bound for $\mathcal{P}_u(r_u < \lambda)$. This is equivalent to $\mathcal{P}_u(r_u - \mu < \lambda - \mu)$, where $\mu = E(r_u)$. We can use Cantelli inequality if $\mu > \lambda$. From Propositions 4 and 5 we find that

$$\mu = M(M-1)\sigma^2$$

and

$$\sigma^2_{r_u} = 4(M-1)^2 s^2 \sigma^2 + 2M(M-1)\sigma^4$$

and the final result directly follows. ◇

From (31) we can conclude that

1) The upper bound of $\mathcal{P}_o$ is not a function of the SNR;
2) For fixed L, the upper bound of $\mathcal{P}_o$ is a monotonic decreasing function of M; and
3) For fixed M, the upper bound of $\mathcal{P}_o$ is a monotonic increasing function of L. (The minimum is achieved when $L = 1$.)

When $\eta > \frac{\ln \frac{M}{L}}{M-L}$, from (33) we deduce that

1) The upper bound of $\mathcal{P}_u$ is a function of the partial SNR, η, and it decreases as η increases,
2) For fixed L, the upper bound of $\mathcal{P}_u$ is a monotonic decreasing function of M; and
3) For fixed M, the upper bound of $\mathcal{P}_u$ is a monotonic decreasing function of L. (The minimum is achieved when $L = M - 1$.)

In Figs. 1 and 2 we show the Cantelli upper bound of $\mathcal{P}_o$ as a function of L and M respectively. Likewise, in Figs. 3 and 4 we show the upper bound of $\mathcal{P}_u$. In Fig. 4 several curves are plotted, each corresponding to a different partial SNR. When

$$\eta \gg \frac{\ln \frac{M}{L}}{M-L}$$

then L basically does not affect the upper bound of $\mathcal{P}_u$. It can easily be shown that this upper bound becomes negligible when compared to the bound of $\mathcal{P}_o$. From the example in Fig. 4, when $L > 120$, the bound becames smaller than 1×10^{-3}.

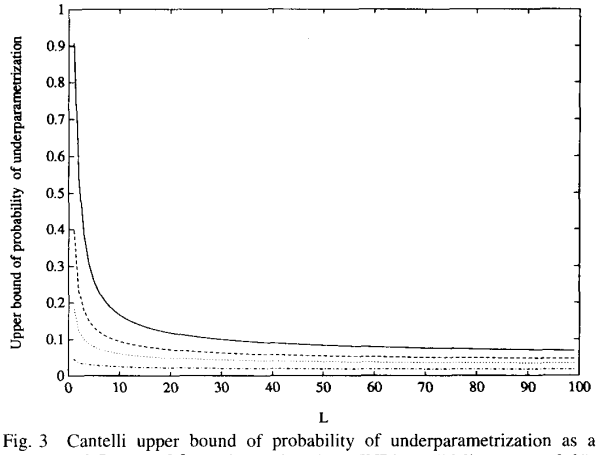

Fig. 3 Cantelli upper bound of probability of underparametrization as a function of L when $M = 100$, and various SNR's (solid line: $\eta_1 = 0.05$; dashed line: $\eta_2 = 0.06$; dotted line: $\eta_3 = 0.07$; dash-dot line: $\eta_4 = 0.1$).

In Fig. 5 we show the upper bound of the probability of error as a function of L for several partial SNR's. It was obtained by adding the upper bounds of over- and underparametrization.

The analysis of the bound of $\mathcal{P}_\epsilon$ entails the following result. When the partial SNR is greater than a particular threshold, γ, the best choice of L according to the upper bound is $L = 1$. γ itself as a function of M decreases when M grows. As the SNR decreases and is below γ, the optimal L that minimizes the upper bound of $\mathcal{P}_\epsilon$ grows.

A similar analysis can be carried out for the sequential estimator. It should be noted that there are other approaches that may be tried to find an optimal strategy for estimation-validation. One of them, for example, could be based on examination of distances among the constructed predictive densities. The idea is to find how close these densities are to each other as a function of L and fixed M. Intuitively, we expect that if they are further apart, it will be easier to discriminate the appropriate model. As a familiar measure of "distance" (discrepancy) between two densities we tried the divergence [19]. Although the selection of L according to the

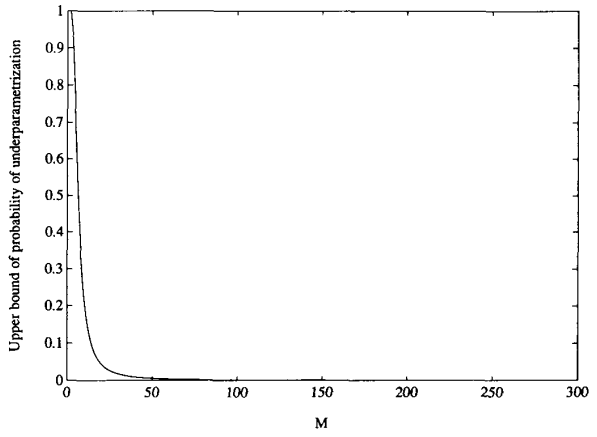

Fig.4 Cantelli upper bound of probability of underparametrization as a function of M when $L = 1$ and SNR $\eta_1 = 0.7$.

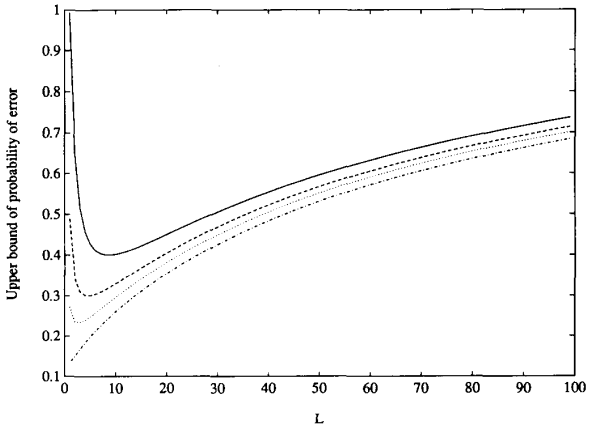

Fig. 5 Upper bound of the probability of error as a function of L when $M = 100$ and various SNR's (solid line: $\eta_1 = 0.05$; dashed line: $\eta_2 = 0.06$; dotted line: $\eta_3 = 0.07$; dash-dot line: $\eta_4 = 0.1$).

maximum divergence, would not imply that the probability of incorrect model selection $\mathcal{P}_\epsilon$ would be minimized [14], it is a reasonable criterion and has been used in the literature. For example, in [2] the divergence was similarly employed to judge which approach was better in estimating a probability density function, the predictive or the estimative. In pattern recognition, for instance, it is a common practice to use the divergence as a class separability measure in feature selection problems when the overall probability of misclassification is of primary interest [12]. In addition, if the divergence can be determined, it can be used to find the upper bound of the probability of error. When two hypotheses are based on Gaussian densities, then [13]

$$\mathcal{P}_\epsilon \leq \frac{1}{2}\left(\frac{\mathcal{J}}{4}\right)^{-\frac{1}{4}}$$

where $\mathcal{J}$ is the divergence between the densities.

The analysis based on divergence will not be presented here. The reason is twofold: 1) it is extensive, and 2) the conclusions are the same as given here. For details, however, the interested reader is referred to [8].

TABLE V
COMPARISON OF PERFORMANCE OF SEQUENTIAL AND SYMMETRIC ESTIMATORS; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE CORRECT MODEL ORDER IS 3

M=50, N=20								
	$\hat{m}_{\text{sym}}(M,L)$				$\hat{m}_{\text{seq}}(M,L)$			
k	$L=1$	$L=15$	$L=25$	$L=49$	$L=1$	$L=15$	$L=25$	$L=49$
1	0	0	0	0	0	0	0	179
2	0	0	0	0	1	8	23	209
3	969	864	819	739	963	753	618	166
4	30	83	100	123	34	125	161	84
5	0	32	42	67	1	58	88	70
6	1	13	19	31	1	30	41	69
7	0	4	10	20	0	12	28	91
8	0	4	10	20	0	14	41	132

We checked the derivations in this section by simulations using several symmetric and sequential estimators. Tha data model was the same as in (15) with the same parameters as in Table I. The vector θ_8 was

$$\theta_8^T = [0.795\ 0.447\ 0.251\ 0\ 0\ 0\ 0\ 0]. \tag{34}$$

There were $M = 50$ data records, each $N = 20$ samples long. Table V shows the results obtained by the various symmetric and sequential estimators in 1000 trials. Best results yielded $\hat{m}_{\text{sym}}(50,1)$ and $\hat{m}_{\text{seq}}(50,1)$. On the other hand, the performance of $\hat{m}_{\text{seq}}(50,49)$ was extremely poor. By extensive simulations we verified that for improved model selection performance it was better to keep the number of estimation sequences small and validation sequences large. When the SNR ratio was very low, as predicted by the analysis, the performance varied with L such that the lower the SNR, the larger the L for the estimator with best performance.

In conclusion, from what seemed to be a subjective choice (the number of estimation sequences), there is not much left. The theoretical arguments and the simulation results suggest that L should be kept small. Although, in a small range of partial SNRs $L = 1$ is not the best choice, it is recommended for use.

VI. RELATION TO OTHER MODEL SELECTION SCHEMES

It will be interesting to compare the model order estimators from this paper to others derived using different principles. Akaike developed a criterion (AIC) exploiting information theoretic arguments. This criterion basically maximizes the expected log-likelihood of a model determined by the method of maximum likelihood [3], and is based on

$$J_k^{(\text{AIC})} = -2\ln f(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M|\hat{\theta}_k) + 2k$$

where $\hat{\theta}_k$ is the maximum likelihood estimate of θ_k, and k is the dimension of the model. Applied to our linear regression problem, this principle will yield the following estimate

$$\hat{m}_{\text{AIC}} = \arg\left\{\min_k \left(-\mathbf{y}^{(M)^T}\mathbf{P}_k\mathbf{y}^{(M)} + 2kM\sigma^2\right)\right\}. \tag{35}$$

It is interesting to note that it is identical to Mallows' C_p conditional mean square error prediction criterion [23].

Rissanen [25] and Schwarz [28] came up with a selection rule with identical functional form using different approaches. Schwarz took the Bayesian route using asymptotic expansion of the posterior probability of the model,[5] while Rissanen set up the problem as a minimization of the bit representation of a signal under different models (minimum description length (MDL) criterion.) Their selection rule rests on [34]

$$J_k^{(\mathrm{MDL})} = -\ln f(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M | \hat{\theta}_k) + \frac{k}{2} \ln M.$$

When this criterion is applied to our problem, it yields

$$\hat{m}_{\mathrm{MDL}} = \arg \left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + kM\sigma^2 \ln M \right) \right\}. \tag{36}$$

Now we shall establish an interesting asymptotic relationship between the AIC and the symmetric estimator. If $L = M - 1$, the symmetric estimator (20) becomes

$$\hat{m}_{\mathrm{sym}}(M, M-1) = \arg \left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k \mathbf{y}_j + kM(M-1)\sigma^2 \ln \frac{M}{M-1} \right) \right\}. \tag{37}$$

For large M

$$\ln \frac{M}{M-1} = \ln \left(1 + \frac{1}{M-1}\right) \simeq \frac{1}{M-1}$$

which implies that approximately

$$\hat{m}(M, M-1) = \arg \left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \sum_{i=1}^{M} \mathbf{y}_i^T \mathbf{P}_k \mathbf{y}_i + kM\sigma^2 \right) \right\}.$$

Next we want to substitute the second term in (37) by its expected value due to overparametrization. For two consecutive models, $\mathcal{M}_{k+1}$ and $\mathcal{M}_k$, and $k \geq m$, we can write

$$\sum_{j=1}^{M} \mathbf{y}_j^T (\mathbf{P}_{k+1} - \mathbf{P}_k) \mathbf{y}_j = \sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P} \mathbf{y}_j$$

where $\mathbf{P}$ is a projection matrix defined by

$$\mathbf{P} = \mathbf{P}_{k+1} - \mathbf{P}_k.$$

Note that it is related to $\mathbf{P}_k$ by

$$\mathbf{P}\mathbf{P}_k = \mathbf{0}.$$

Since

$$\theta_m^T \mathbf{H}_m^T \mathbf{P} \mathbf{H}_m \theta_m = 0$$

we have

$$\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P} \mathbf{y}_j = \sum_{j=1}^{M} \mathbf{e}_j^T \mathbf{P} \mathbf{e}_j > 0.$$

[5]It should be pointed out that Kashyap [16] and Leonard [21] did a somewhat more general derivation than Schwarz.

The distribution of $\frac{\mathbf{e}_j^T \mathbf{P} \mathbf{e}_j}{\sigma^2}$ is χ_1^2. Thus

$$E\left(\sum_{j=1}^{M} \mathbf{e}_j^T \mathbf{P} \mathbf{e}_j \right) = M\sigma^2.$$

This is the expected value of the difference in "penalization" induced by $\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k \mathbf{y}_j$ between two successive models for $k \geq m$. Going back to (37) and substituting the term $\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k \mathbf{y}_j$ by the expected penalization, we find that

$$\hat{\hat{m}}_{\mathrm{sym}}(M, M-1) = \arg \left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + 2kM\sigma^2 \right) \right\}.$$

This is *identical* to (35), the AIC estimator. Consequently, for large M we would expect that $\hat{m}_{\mathrm{sym}}(M, M-1)$ and $\hat{m}_{\mathrm{AIC}}$ should yield statistically comparable results. When the two estimators were compared by extensive simulations, this was indeed the case. This result is analogous to Stone's [32], where he showed that asymptotically the cross-validation criterion and the AIC were equivalent. Note, however, that our setting is different from that in [32].

If in (20) and (21) we substitute $L = 1$, we find that

$$\hat{m}_{\mathrm{sym}}(M, 1) = \arg \left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k \mathbf{y}_j + kM\sigma^2 \ln M \right) \right\} \tag{38}$$

and

$$\hat{m}_{\mathrm{seq}}(M, 1) = \arg \left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + M\mathbf{y}_1^T \mathbf{P}_k \mathbf{y}_1 + kM\sigma^2 \ln M \right) \right\}. \tag{39}$$

Since the penalty for overparametrization obtained from the second terms in (38) and (39) is always positive (on average $kM\sigma^2$), it is clear that the overall penalties of these estimators will be more stringent than the penalty of the MDL estimator. We expect, therefore, that the lowest order model will be chosen more likely by $\hat{m}_{\mathrm{sym}}(M, 1)$ and $\hat{m}_{\mathrm{seq}}(M, 1)$ than by the MDL estimator. Moreover, from a comparison of (38) and (39) we would expect that the symmetric and the sequential estimator will have similar performance.

In Tables VI–VIII, simulation results are presented that illustrate the performance achieved by all these estimators. The same model was used as in (14) with parameters given in Table I and (34). The number of sequences M and L was varied. Best results were obtained by $\hat{m}_{\mathrm{sym}}(M, 1)$. It outperformed the AIC and MDL rules in all the simulations. Note also that, as anticipated, the difference between the $\hat{m}_{\mathrm{AIC}}$ and $\hat{m}_{\mathrm{sym}}(M, M-1)$ was statistically insignificant. In addition, $\hat{m}_{\mathrm{seq}}(M, 1)$ performed almost as well as $\hat{m}_{\mathrm{sym}}(M, 1)$. The results achieved by $\hat{m}_{\mathrm{seq}}(M, M-1)$ were completely unreliable.

TABLE VI
COMPARISON AMONG ESTIMATORS; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE CORRECT MODEL ORDER IS 3

$M = 10, N = 20$

k	$\hat{m}_{\text{sym}}(10,1)$	$\hat{m}_{\text{sym}}(10,5)$	$\hat{m}_{\text{sym}}(10,9)$	$\hat{m}_{\text{seq}}(10,1)$	$\hat{m}_{\text{seq}}(10,5)$	$\hat{m}_{\text{seq}}(10,9)$	$\hat{m}_{\text{AIC}}$	$\hat{m}_{\text{MDL}}$
1	1	0	0	7	19	145	0	0
2	284	177	138	312	205	226	135	167
3	643	627	603	587	477	196	607	631
4	46	92	97	59	109	104	100	89
5	21	53	78	22	73	92	71	53
6	2	21	37	7	51	66	34	24
7	2	18	24	5	39	70	30	24
8	1	12	23	1	27	101	23	12

TABLE VII
COMPARISON AMONG ESTIMATORS; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE CORRECT MODEL ORDER IS 3

$M = 40, N = 20$

k	$\hat{m}_{\text{sym}}(40,1)$	$\hat{m}_{\text{sym}}(40,20)$	$\hat{m}_{\text{sym}}(40,39)$	$\hat{m}_{\text{seq}}(40,1)$	$\hat{m}_{\text{seq}}(40,20)$	$\hat{m}_{\text{seq}}(40,39)$	$\hat{m}_{\text{AIC}}$	$\hat{m}_{\text{MDL}}$
1	0	0	0	0	0	181	0	0
2	2	0	0	3	31	210	0	0
3	967	811	746	950	630	165	743	928
4	24	93	108	38	120	93	107	49
5	5	51	58	8	77	82	63	17
6	2	26	39	1	62	64	38	5
7	0	11	25	0	41	80	25	1
8	0	8	24	0	39	125	24	0

TABLE VIII
COMPARISON AMONG ESTIMATORS; THE VALUES SHOW THE NUMBER OF TIMES THE MODEL ORDER k WAS CHOSEN IN 1000 REALIZATIONS BY THE RESPECTIVE ESTIMATORS; THE CORRECT MODEL ORDER IS 3

$M = 40, N = 40$

k	$\hat{m}_{\text{sym}}(40,1)$	$\hat{m}_{\text{sym}}(40,20)$	$\hat{m}_{\text{sym}}(40,39)$	$\hat{m}_{\text{seq}}(40,1)$	$\hat{m}_{\text{seq}}(40,20)$	$\hat{m}_{\text{seq}}(40,39)$	$\hat{m}_{\text{AIC}}$	$\hat{m}_{\text{MDL}}$
1	0	0	0	0	0	111	0	0
2	0	0	0	0	3	192	0	0
3	970	812	736	962	666	204	742	933
4	24	93	113	29	126	88	106	48
5	4	45	67	6	65	88	62	14
6	1	25	37	2	45	87	43	3
7	1	15	26	0	50	84	25	1
8	0	10	21	1	45	146	22	1

VII. CONCLUSION

In this paper, model (order) selection criteria were derived based on Bayesian predictive densities and multiple data records. In their derivation, the underlying principle was to measure the models' performances *only* by data which were not used for their estimation. Several important issues were addressed, such as consistency and choice of estimation and validation data records. It was proved that the selection rules are consistent when the set of data records for estimation is fixed and the number of data records for validation tends to infinity. On the contrary, when the set of validation data records is fixed, and the number of estimation data records tends to infinity, the rules are inconsistent. In addition, it was shown that the probability of overparametrization is minimized when the number of estimation data records is equal to one. On the other hand, the probability of underparametrization is minimized when the number of validation records is equal to one. Upper bounds of these probabilities are derived. These bounds suggest that it is better to keep the number of estimation data records low. The asymptotical analysis shows that the Bayes selection rule becomes equivalent to AIC if only one data record is used for validation. In addition, if one data record is used for estimation, the Bayes rule has a more stringent penalty than the MDL. Extensive simulation results are presented. They support the theoretical analysis in the paper. Moreover, they show that the Bayesian selection rules have better performance than the AIC and MDL criteria.

To allow mathematical tractability and insight into the problem, we analyzed a set of nested linear models in a fairly restrictive scenario. Most of these restrictions, however, can be removed, and selection rules can be derived along the same lines for more realistic cases. This is possible due to the coherency of the Bayesian theory. For instance, the case of nested linear models and *unknown* σ^2 can be handled readily (see [8]). Selections from more complex sets of models will be presented in a follow-up paper. It should also be clear that the same idea can be used to derive selection rules for the most often encountered case in practice—when only one data record is observed. If the data sequence is segmented in M disjoint subsequences, we are back to the multiple-data-record case [9].

APPENDIX A

Derivation of the General Symmetric and Sequential Estimators: First the symmetric estimator is derived. Suppose that we have two independent sequences with lengths N_1 and N_2,

generated by

$$\mathbf{y}_1 = \mathbf{H}_{N_1 m}\theta_m + \mathbf{e}_1 \tag{A-1}$$

and

$$\mathbf{y}_2 = \mathbf{H}_{N_2 m}\theta_m + \mathbf{e}_2 \tag{A-2}$$

where $\mathbf{H}_{N_1 m}$ and $\mathbf{H}_{N_2 m}$ are $N_1 \times m$ and $N_2 \times m$ matrices, respectively, whose ranks are equal to m. (Note that we use $N_1 m$ and $N_2 m$ as indices of $\mathbf{H}$ to emphasize that the dimensions of the two matrices are not necessarily identical as they were before. If it is clear that the $\mathbf{H}$ matrices are identical, we shall use as an index the number of their columns only.)

Let the parameter vectors θ_m in (A-1) and (A-2) be identical. Now if we write the predictive density of $\mathbf{y}_2$ according to $\mathbf{y}_1$ and the model $\mathcal{M}_k$ as

$$f(\mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k) = \int_{\Theta_k} f(\mathbf{y}_2|\theta_k, \mathbf{y}_1, \mathcal{M}_k) f(\theta_k|\mathbf{y}_1, \mathcal{M}_k) d\theta_k$$

we obtain

$$f(\mathbf{y}_2|\mathbf{y}_1, \mathcal{M}_k) \propto \frac{|\mathbf{H}_{N_1 k}^T \mathbf{H}_{N_1 k}|^{\frac{1}{2}}}{|\tilde{\mathbf{H}}_k^T \tilde{\mathbf{H}}_k|^{\frac{1}{2}}}$$

$$\cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\tilde{\mathbf{y}}^T \tilde{\mathbf{P}}_k^{\perp} \tilde{\mathbf{y}} - \mathbf{y}_1^T \mathbf{P}_{N_1 k}^{\perp} \mathbf{y}_1\right)\right\} \tag{A-3}$$

where

$$\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \qquad \tilde{\mathbf{H}}_k = \begin{pmatrix} \mathbf{H}_{N_1 k} \\ \mathbf{H}_{N_2 k} \end{pmatrix}$$

and

$$\tilde{\mathbf{P}}_k^{\perp} = \mathbf{I} - \tilde{\mathbf{H}}_k \left(\tilde{\mathbf{H}}_k^T \tilde{\mathbf{H}}_k\right)^{-1} \tilde{\mathbf{H}}_k^T$$

$$\mathbf{P}_{N_1 k}^{\perp} = \mathbf{I} - \mathbf{H}_{N_1 k} \left(\mathbf{H}_{N_1 k}^T \mathbf{H}_{N_1 k}\right)^{-1} \mathbf{H}_{N_1 k}^T.$$

Now, suppose there are M sequences, each N samples long. If L of them are used for estimation, then using (A-3)

$$f(\tilde{\mathbf{y}}_{M-L}|\tilde{\mathbf{y}}_L, \mathcal{M})$$

$$\propto \frac{|\tilde{\mathbf{H}}_{Lk}^T \tilde{\mathbf{H}}_{Lk}|^{\frac{1}{2}}}{|\tilde{\mathbf{H}}_{Mk}^T \tilde{\mathbf{H}}_{Mk}|^{\frac{1}{2}}} \exp\{-\frac{1}{2\rho^2}(\tilde{\mathbf{y}}_M^T \tilde{\mathbf{P}}_{Mk}^{\perp} \tilde{\mathbf{y}}_M - \tilde{\mathbf{y}}_L^T \tilde{\mathbf{P}}_{Lk}^{\perp} \tilde{\mathbf{y}}_L)\}$$

where $\tilde{\mathbf{y}}_{M-L}$ and $\tilde{\mathbf{y}}_L$ are formed by concatenating the validation and estimation sequences respectively. Similarly, $\tilde{\mathbf{y}}_M$ represents all the sequences $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M$ stacked in one vector. The matrices $\tilde{\mathbf{H}}_{Mk}$ and $\tilde{\mathbf{H}}_{Lk}$ are special $M \times 1$ and $L \times 1$ block matrices whose blocks are identical to $\mathbf{H}_k$. The projection matrices $\tilde{\mathbf{P}}_{Mk}^{\perp}$ and $\tilde{\mathbf{P}}_{Lk}^{\perp}$ are then $M \times M$ and $L \times L$ block matrices respectively, whose diagonal blocks are equal to $\mathbf{I} - \frac{1}{M}\mathbf{P}_k$ and $\mathbf{I} - \frac{1}{L}\mathbf{P}_k$ respectively, and off-diagonal blocks are $\frac{1}{M}\mathbf{P}_k$ and $\frac{1}{L}\mathbf{P}_k$ respectively.

Next, note that

$$|\tilde{\mathbf{H}}_{Mk}^T \tilde{\mathbf{H}}_{Mk}| = |M \mathbf{H}_k^T \mathbf{H}_k|$$

and

$$|\tilde{\mathbf{H}}_{Lk}^T \tilde{\mathbf{H}}_{Lk}| = |L \mathbf{H}_k^T \mathbf{H}_k|.$$

Thus,

$$\frac{|\tilde{\mathbf{H}}_{Lk}^T \tilde{\mathbf{H}}_{Lk}|^{\frac{1}{2}}}{|\tilde{\mathbf{H}}_{Mk}^T \tilde{\mathbf{H}}_{Mk}|^{\frac{1}{2}}} = \left(\frac{L}{M}\right)^{\frac{k}{2}}.$$

We can also show that

$$\tilde{\mathbf{y}}_M^T \tilde{\mathbf{P}}_{Mk}^{\perp} \tilde{\mathbf{y}}_M = \tilde{\mathbf{y}}_M^T \tilde{\mathbf{y}}_M - \frac{1}{M} \mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)}$$

where

$$\mathbf{y}^{(M)} = \sum_{i=1}^{M} \mathbf{y}_i.$$

Moreover,

$$\tilde{\mathbf{y}}_L^T \tilde{\mathbf{P}}_{Lk}^{\perp} \tilde{\mathbf{y}}_L = \tilde{\mathbf{y}}_L^T \tilde{\mathbf{y}}_L - \frac{1}{L} \mathbf{y}^{(L)^T} \mathbf{P}_k \mathbf{y}^{(L)}$$

where

$$\mathbf{y}^{(L)} = \sum_{i=1}^{L} \mathbf{y}_i.$$

Now, the predictive density from (A-4) can be rewritten as

$$\ln f(\tilde{\mathbf{y}}_{M-L}|\tilde{\mathbf{y}}_L, \mathcal{M}_k) = C + k\sigma^2 \ln \frac{M}{L} + \frac{1}{M} \mathbf{y}^{(M)^T} \cdot \mathbf{P}_k \mathbf{y}^{(M)} - \frac{1}{L} \mathbf{y}^{(L)^T} \mathbf{P}_k \mathbf{y}^{(L)} \tag{A-5}$$

where C is a constant independent of k.

We are interested in determining the number of crosscorrelation terms $\mathbf{y}_i^T \mathbf{P}_k \mathbf{y}_j$ for fixed i and j in (A-5). Since there are C_M^L combinations of choosing the estimation and validation sets of data records, where

$$C_M^L = \frac{M!}{(M-L)!L!},$$

the term $\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)}$ will yield the crosscorrelation $\mathbf{y}_i^T \mathbf{P}_k \mathbf{y}_j$, C_M^L times. The same crosscorrelation product will occur in $\mathbf{y}^{(L)^T} \mathbf{P}_k \mathbf{y}^{(L)}$, C_{M-2}^{L-2} times for $L \geq 2$. After the symmetrization, all the terms of the form $\mathbf{y}_i^T \mathbf{P}_k \mathbf{y}_i$ will drop out. Therefore, the selection rule will rest on

$$J_k'(M, L) = -\left(\frac{1}{M}\binom{M}{L} - \frac{1}{L}\binom{M-2}{L-2}\right) \cdot \sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k (\mathbf{y}^{(M)} - \mathbf{y}_j) + k\binom{M}{L}\sigma^2 \ln \frac{M}{L}.$$

This may be simplified to

$$J_k(M, L) = -\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k (\mathbf{y}^{(M)} - \mathbf{y}_j) + k\frac{M(M-1)}{M-L}\sigma^2 \ln \frac{M}{L}.$$

Since the estimator chooses the model with minimum $J_k(M, L)$, we finally obtain

$$\hat{m}_{\text{sym}}(M, L) = \arg\left\{\min_k \left(-\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k (\mathbf{y}^{(M)} - \mathbf{y}_j) + k\frac{M(M-1)}{M-L}\sigma^2 \ln \frac{M}{L}\right)\right\}$$

with

$$\mathbf{y}^{(M)} = \sum_{i=1}^{M} \mathbf{y}_j.$$

The general form of the sequential estimator is easily found from (A-4). It is

$$\hat{m}_{\text{seq}}(M, L) = \arg\left\{ \min_k \left(-\mathbf{y}^{(M)^T} \mathbf{P}_k \mathbf{y}^{(M)} + \frac{M}{L} \mathbf{y}^{(L)^T} \mathbf{P}_k \mathbf{y}^{(L)} + kM\sigma^2 \ln \frac{M}{L} \right) \right\}.$$

APPENDIX B

Proof of Proposition 1: Let

$$J_k = -\sum_{j=1}^{M} \mathbf{y}_j^T \mathbf{P}_k (\mathbf{y}^{(M)} - \mathbf{y}_j) + k \frac{M(M-1)}{M-L} \sigma^2 \ln \frac{M}{L}$$

where

$$\mathbf{y}^{(M)} = \sum_{j=1}^{M} \mathbf{y}_j.$$

Note that

$$1 \le L \le M - 1.$$

Define

$$\Delta J_k = J_k(M, L) - J_m(M, L). \tag{B-1}$$

We will show that

$$p\lim \frac{\Delta J_k}{F(M)} > 0 \quad \text{for } k \neq m$$

where $F(M)$ is a suitably chosen function of M such that $F(M) > 0$, which implies that the symmetric estimator is consistent, since J_k achieves a minimum for $k = m$. ($p\lim \frac{\Delta J_k}{F(M)} > 0$ denotes $\lim_{M\to\infty} \mathcal{P}[|\frac{\Delta J_k}{F(M)} - c| > \delta] = 0$ and $c > 0$).

Suppose first that $k < m$. Then (B-1) yields

$$\Delta J_k = (k-m) \frac{M(M-1)}{M-L} \sigma^2 \ln \frac{M}{L} - \sum_{j=1}^{M} \mathbf{y}_j^T (\mathbf{P}_k - \mathbf{P}_m)(\mathbf{y}^{(M)} - \mathbf{y}_j).$$

We may write

$$\mathbf{P}_m = \mathbf{P}_k + \mathbf{P}_{mk}^{\perp}$$

where $\mathbf{P}_{mk}^{\perp}$ is a projection matrix of rank $m-k$ and $\mathbf{P}_k \mathbf{P}_{mk}^{\perp} = \mathbf{0}$. Then

$$-\sum_{j=1}^{M} \mathbf{y}_j^T (\mathbf{P}_k - \mathbf{P}_m)(\mathbf{y}^{(M)} - \mathbf{y}_j) = \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbf{y}_i^T \mathbf{P}_{mk}^{\perp} \mathbf{y}_j.$$

Since

$$\mathbf{y}_j = \mathbf{H}_m \theta_m + \mathbf{e}_j$$

we have

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbf{y}_i^T \mathbf{P}_{mk}^{\perp} \mathbf{y}_j = M(M-1) \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \mathbf{H}_m \theta_m + 2(M-1) \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \sum_{j=1}^{M} \mathbf{e}_j + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbf{e}_i^T \mathbf{P}_{mk}^{\perp} \mathbf{e}_j.$$

Thus

$$\Delta J_k = (k-m) \frac{M(M-1)}{M-L} \sigma^2 \ln \frac{M}{L} + M(M-1) \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \mathbf{H}_m \theta_m + 2(M-1) \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \sum_{j=1}^{M} \mathbf{e}_j + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbf{e}_i^T \mathbf{P}_{mk}^{\perp} \mathbf{e}_j.$$

Now we shall show that

$$p\lim \frac{1}{M^2} \Delta J_k > 0.$$

Using Slutsky's theorem

$$\begin{aligned} p\lim \frac{1}{M^2} \Delta J_k &= p\lim \left(\frac{(k-m)M(M-1)}{M^2(M-L)} \sigma^2 \ln \frac{M}{L} + \frac{M(M-1)}{M^2} \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \mathbf{H}_m \theta_m + \frac{2(M-1)}{M^2} \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \sum_{j=1}^{M} \mathbf{e}_j + \frac{1}{M^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbf{e}_i^T \mathbf{P}_{mk}^{\perp} \mathbf{e}_j \right) \\ &= p\lim \left(\frac{(k-m)M(M-1)}{M^2(M-L)} \sigma^2 \ln \frac{M}{L} \right) + p\lim \left(\frac{M(M-1)}{M^2} \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \mathbf{H}_m \theta_m \right) + p\lim \left(\frac{2(M-1)}{M^2} \theta_m^T \mathbf{H}_m^T \mathbf{P}_{mk}^{\perp} \sum_{j=1}^{M} \mathbf{e}_j \right) + p\lim \left(\frac{1}{M^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbf{e}_i^T \mathbf{P}_{mk}^{\perp} \mathbf{e}_j \right). \end{aligned} \tag{B-2}$$

It is obvious that

$$p\lim \left(\frac{(k-m)M(M-1)}{M^2(M-L)} \sigma^2 \ln \frac{M}{L} \right) = \lim_{M\to\infty} \left(\frac{(k-m)M(M-1)}{M^2(M-L)} \sigma^2 \ln \frac{M}{L} \right) = 0 \tag{B-3}$$

and

$$\begin{aligned} p\lim\left(\frac{M(M-1)}{M^2}\theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\mathbf{H}_m\theta_m\right) \\ = \lim_{M\to\infty}\left(\frac{M(M-1)}{M^2}\theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\mathbf{H}_m\theta_m\right) \\ = \theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\mathbf{H}_m\theta_m = c > 0. \end{aligned} \tag{B-4}$$

Next

$$p\lim\left(\frac{2(M-1)}{M^2}\theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\sum_{j=1}^{M}\mathbf{e}_j\right) = 0 \tag{B-5}$$

since

$$\theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\sum_{j=1}^{M}\mathbf{e}_j = O_p(M^{\frac{1}{2}}) \tag{B-6}$$

where $O_p(M^{\frac{1}{2}})$ denotes that the sequence of random variables $r_M = \theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\sum_{j=1}^{M}\mathbf{e}_j$ is at most of order in probability $M^{\frac{1}{2}}$. To show this, note that

$$E(r_M) = 0$$

and

$$E(r_M^2) = M\sigma^2 N_1$$

where $N_1 = \theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\mathbf{H}_m\theta_m$. Applying Chebyshev's inequality

$$\mathcal{P}[|r_M| \geq M^{\frac{1}{2}}N_2] \leq \frac{M\sigma^2 N_1}{MN_2^2}.$$

If for a given δ we choose N_2 such that $\frac{\sigma^2 N_1}{N_2^2} \leq \delta$, then

$$\mathcal{P}[M^{-\frac{1}{2}}|r_M| \geq N_2] \leq \delta$$

where δ may be arbitrarily small. The last expression is equivalent to (B-6).

Finally

$$p\lim\frac{1}{M^2}\sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathbf{e}_i^T\mathbf{P}_{mk}^{\perp}\mathbf{e}_j = 0. \tag{B-7}$$

We shall show that $\sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathbf{e}_i^T\mathbf{P}_{mk}^{\perp}\mathbf{e}_j = O_p(M)$, which will imply (B-7). Define the random variable v as

$$v = \sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathbf{e}_i^T\mathbf{P}_{mk}^{\perp}\mathbf{e}_j. \tag{B-8}$$

Its mean value and variance may readily be found using Propositions 4 and 5. They are given by

$$E(v) = 0. \tag{B-9}$$

and

$$E(v - E(v))^2 = E(v^2) = 2(m-k)M(M-1)\sigma^4.$$

Applying Chebyshev's inequality,

$$\mathcal{P}[|v \geq MN_2|] \leq \frac{2M(M-1)\sigma^4(m-k)}{M^2N_2^2}$$

which implies that $v = O_p(M)$, and therefore (B-7).

Using (B-3), (B-4), (B-5), and (B-7) in (B-2) yields

$$p\lim\frac{1}{M^2}\Delta J_k(M,L) = \theta_m^T\mathbf{H}_m^T\mathbf{P}_{mk}^{\perp}\mathbf{H}_m\theta_m = c > 0. \tag{B-10}$$

Now, let $k > m$. Then

$$\Delta J_k = (k-m)\frac{M(M-1)}{M-L}\sigma^2\ln\frac{M}{L} - \sum_{j=1}^{M}\mathbf{y}_j^T(\mathbf{P}_k - \mathbf{P}_m)(\mathbf{y} - \mathbf{y}_j).$$

Since $k > m$

$$\mathbf{P}_k = \mathbf{P}_m + \mathbf{P}_{km}^{\perp}.$$

Furthermore

$$\mathbf{P}_{km}^{\perp}\mathbf{H}_m = \mathbf{0}.$$

Therefore

$$\Delta J_k = (k-m)\frac{M(M-1)}{M-L}\sigma^2\ln\frac{M}{L} - \sum_{j=1}^{M}\mathbf{e}_j^T\mathbf{P}_{km}^{\perp}(\mathbf{e} - \mathbf{e}_j).$$

$$= (k-m)\frac{M(M-1)}{M-L}\sigma^2\ln\frac{M}{L} - \sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathbf{e}_i^T\mathbf{P}_{km}^{\perp}\mathbf{e}_j. \tag{B-11}$$

Now, it immediately follows that

$$p\lim\frac{\Delta J_k}{(k-m)M\sigma^2\ln\frac{M}{L}} = 1 \tag{B-12}$$

since it can be shown similarly as for $\sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathbf{e}_i^T\mathbf{P}_{mk}^{\perp}\mathbf{e}_j$ that $\sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathbf{e}_i^T\mathbf{P}_{km}^{\perp}\mathbf{e}_j = O_p(M)$.

From (B-10) and (B-12) we may assert that the estimator (20) is consistent.

REFERENCES

[1] J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis*. New York: Cambridge University Press, 1975.

[2] J. Aitchison, "Goodness of prediction fit," *Biometrika*, vol. 62, pp. 547–554, 1975.

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

[4] A. C. Atkinson, "Posterior probabilities for choosing a regression model," *Biometrika*, vol. 65, pp. 39–48, 1978.

[5] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, 1973.

[6] H. Clergeot, "Filter-order selection in adaptive maximum likelihood estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 199–210, 1984.

[7] E. J. Delp, R. L. Kashyap, and O. R. Mitchell, "Image data compression using autoregressive time series models," *Pattern Recognition*, vol. 11, pp. 313–323, 1979.

[8] P. M. Djurić, "Selection of signal and system models by Bayesian predictive densities," Ph.D. dissertation, Univ. of Rhode Island, Kingston, 1990.

[9] P. M. Djurić, "Model selection by cross-validation," in *Proc. IEEE ISCAS* (New Orleans, LA), 1990, pp. 2254–2257.

[10] P. M. Djurić and S. M. Kay, "Predictive probability as a criterion for model selection," in *Proc. IEEE ICASSP* (Albuquerque, NM), 1990, pp. 2415–2418.
[11] S. Geisser, "The predictive sample reuse method with applications," *J. Amer. Statist. Assoc.*, vol. 70, pp. 320–328, 1975.
[12] D. J. Hand, *Discrimination and Classification*. New York: Wiley, 1981.
[13] T. T. Kadota and L. L. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 278–284, 1967.
[14] T. Kailath, "The divergence and Bhatacharya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, pp. 52–60, 1967.
[15] R. L. Kashyap, "A Bayesian comparison of different classes of dynamic models using empirical data," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 715–727, 1977.
[16] R. L. Kashyap, "Optimal choice of AR and MA parts in autoregressive moving average models," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 4, pp. 99–104, 1982.
[17] M. Kaveh, H. Wang, and H. Hung, "On the theoretical performance of a class of estimators of the number of narrow-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, 1987.
[18] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 1. New York: Hafner, 1969.
[19] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
[20] R. Kumaresan, D. W. Tufts, and L. L. Scharf, "A Prony method for noisy data: Choosing the signal components and selecting the order in exponential signal models," *Proc. IEEE*, vol. 72, pp. 230–233, 1984.
[21] T. Leonard, "Comment" on a paper by M. Lejeune and G. D. Faulkenberry, *J. Am. Statist. Assoc.*, vol. 77, pp. 657–658, 1982.
[22] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, 1975.
[23] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol.15, pp. 661–675, 1973.
[24] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: Wiley, 1973.
[25] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–478, 1978.
[26] H. V. Roberts, "Probabilistic prediction," *J. Amer. Statist. Assoc.*, vol. 60, pp. 50–61, 1965.
[27] U. E. Ruttiman and H. V. Pipberger, "Compression of the ECG by prediction or interpolation and entropy encoding," *IEEE Trans. Biomed. Eng.*, vol. BME-26, pp. 613–623, 1979.
[28] G. Schwarz, "Estimating the dimension of the model," *Annals Statist.*, vol. 6, pp. 461–464, 1978.
[29] S. L. Sclove, "Application of model-selection criteria to some problems in multivariate analysis," *Psychometrika*, vol. 52, pp. 333–343, 1987.
[30] S. R. Searle, *Linear Models*. New York: Wiley, 1971.
[31] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Royal Statist. Soc. (Series B)*, vol. 41, pp. 111–147, 1974.
[32] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *J. Royal Statist. Soc. B*, vol. 39, pp. 44–47, 1977.
[33] H. Wang and M. Kaveh, "On the performance of signal subspace processing—Part I: Narrow-band systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, 1986.
[34] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 387–392, 1985.
[35] Q. Zhang, K. M. Wong, P. C. Yip, and J. P. Reilly, "Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, 1989.

Petar Djurić (S'86–M'90) was born in Strumica, Yugoslavia, in 1957. He received the B.S. and M.S. degrees from the University of Belgrade, Yugoslavia, in 1981 and 1986, respectively, and the Ph.D. degree from the University of Rhode Island, Kingston, RI, in 1990, all in electrical engineering.

Currently, Dr. Djurić is an Assistant Professor in the Department of Electrical Engineering, State University of New York at Stony Brook, NY. His main research interests are in statistical signal processing and system modeling.

Steven Kay (S'76–M'78–SM'83–F'89) was born in Newark, NJ, on April 5, 1951. He received the B.E. degree from Stevens Institute of Technology, Hoboken, NJ, in 1972, the M.S. degree from Columbia University, New York, NY, in 1973, and the Ph.D. degree from Georgia Institute of Technology, Atlanta, GA, in 1980, all in electrical engineering.

From 1972 to 1975 he was with Bell Laboratories, Holmdel, NJ, where he was involved with transmission planning for speech communications and simulation and subjective testing of speech processing algorithms. From 1975 to 1977 he attended Georgia Institute of Technology to study communication theory and digital signal processing. From 1977 to 1980 he was with the Submarine Signal Division, Raytheon, Portsmouth, RI, where he engaged in research on autoregressive spectral estimation and the design of sonar systems. He is presently Professor of Electrical Engineering at the University of Rhode Island, Kingston, and a consultant to industry and the Navy. He has written numerous papers, many of which have been reprinted in the IEEE Press book *Modern Spectrum Analysis II*. He is a contribu``````tor to several edited books on spectral estimation and is the author of *Modern Spectral Estimation: Theory and Application* (Prentice-Hall, 1988) and *Fundamentals of Statistical Signal Processing: Estimation Theory* (Prentice-Hall, 1993). His current interests are spectrum analysis, detection and estimation theory, and statistical signal processing.

Dr. Kay is a member of Tau Beta Pi and Sigma Xi. He has served on the Acoustics, Speech, and Signal Processing Committee on Spectral Estimation and Modeling.