Conditional Model Order Estimation

Steven Kay, *Fellow, IEEE*

***Abstract*—A new approach to model order selection is proposed. Based on the theory of sufficient statistics, the method does not require any prior knowledge of the model parameters. It is able to discriminate between models by basing the decision on the part of the data that is independent of the model parameters. This is accomplished conceptually by transforming the data into a sufficient statistic and an ancillary statistic with respect to the model parameters. It is the probability density function of the ancillary statistic when adjusted for its dimensionality that is used to estimate the order. Furthermore, the rule is directly tied to the goal of minimizing the probability of error and does not employ any asymptotic approximations. The estimator can be shown to be consistent and, via computer simulation, is found to outperform the minimum description length estimator.**

***Index Terms*—Adaptive signal detection, modeling, spectral analysis, speech analysis.**

I. Introduction

THE DETERMINATION of the number of parameters in a model used to fit a data set is a well-known and well-researched problem [1], [2], [11], [16], [17]. When the parameters of each competing model are known, the optimal solution is to choose the model with the largest *a posteriori* probability. However, as is usually the case, the parameters are unknown, and the fitting problem becomes quite difficult with no optimal solution existing. In terms of hypothesis testing, model selection is a multiple composite hypothesis test [15]. One way to attack this problem is to assume a prior probability density function (PDF) for the unknown parameters of each model. This is the so-called Bayesian approach [5], [17]. Then, the model parameters are "integrated out" to yield the PDF of the data. Finally, the maximum *a posteriori* (MAP) decision rule, which minimizes the probability of error, is used. In practice, prior PDFs are seldom available, leading one to assume a noninformative prior or to discard the prior entirely [7]. This approach, however, is fraught with difficulty as stated by Cox and Hinkley. With reference to prior selection, they state the following [6]:

> "There is no difficulty in principle with [prior selection] but note that with the approximate uniform prior densities [the factors due to the priors] do not cancel, as they would in any calculation 'within' an individual model. … From a subjectivist viewpoint it is in principle possible in any particular application to determine 'your' numerical value for this ratio."

Manuscript received March 24, 2000; revised May 17, 2001. This work was supported by the Office of Naval Research under Contract N66604-98-C-2376. The associate editor coordinating the review of this paper and approving it for publication was Prof. Bjorn Ottersten.

The author is with the Department of Electrical and Computer Engineering, University of Rhode Island, Kingston, RI 02881 USA (e-mail: kay@ele.uri.edu).

Publisher Item Identifier S 1053-587X(01)07048-9.

To avoid the use of priors numerous model selection rules have been formulated based on information theoretic concepts. Some of these are the Akaike information criterion (AIC) [2] and its variants, the final prediction error (FPE) [1], and the minimum description length (MDL) [16]. All these criteria have been justified and/or derived based on *asymptotic arguments*. Their application to finite data records (and, in particular, short data records, which is the case of interest) may not be justified. Furthermore, for finite data records, *these criteria do not directly attempt to minimize the probability of error of a decision*. Hence, there is no direct link to the important goal of selecting the correct model order most of the time.

Our approach is an attempt to alleviate the theoretical and practical shortcomings of the asymptotic approaches previously described. We desire a rule that is based on finite data records and works well in terms of maximizing the probability of a correct decision. To this end, we propose the conditional model estimator (CME). It is applicable to the model selection problem when sufficient statistics exist for the parameters of each competing model. The approach is classical in nature and, therefore, does not require any prior knowledge of the model parameters. Based on the theory of sufficient statistics, *the CME avoids the need for prior parameter knowledge by maximizing the conditional PDF of the data*. The conditioning is done on the sufficient statistics, which are observed directly. One way to view this approach is that it splits the observed data into two parts. One part is the sufficient statistic for the model parameters under each hypothesis, and the remaining data has a PDF that does not depend on the model parameters but only on its dimension. The remaining data is sometimes referred to as an ancillary statistic [15] when making inferences about the unknown model parameters. Since the model parameters are unknown and presumably may take on any values, the sufficient statistic PDF cannot be determined in the absence of prior knowledge. Hence, this part of the data is discarded, leaving only the "ancillary" data. The latter carries no information about the model parameters, assuming a *given* model. However, in the model order selection problem, it appears that the PDF of the ancillary statistic *when adjusted for its dimension* is capable of discriminating between models, without requiring knowledge of the model parameters for each model. It will be shown that this approach has some optimality properties in that it yields the minimum variance unbiased (MVU) estimator of the probability of a correct decision under the correct hypothesis. In addition, it can be shown to produce a consistent estimator. Finally, it should be noted that at least in the case of the Gaussian linear model, which is examined in detail in this paper, the CME yields a similar estimator to that obtained based on Bayesian arguments. Bayesian approaches have been summarized in [9], who have tabulated the various model order selection rules for numerous assumed prior proba-

bilities on the parameters. For finite data records, however, the arbitrariness introduced by the various priors is not shared by the CME approach.

One limitation of the proposed method is that it requires the PDF families to admit a minimal set of sufficient statistics. This is, unfortunately, only the case for certain PDFs [15]. However, the methodology proposed can be extended to the case of approximate sufficient statistics, which always exist. In particular, the maximum likelihood estimator (MLE) is asymptotically a sufficient statistic. Of course, pursuing this line of reasoning again leads us to the large data record case, which we sought to avoid. A later paper will explore this avenue. In addition, it should be noted that for best performance, the sufficient statistic should be minimal in dimension. However, the approach can still be applied for nonminimal sufficient statistics, albeit, with a loss in performance. Finally, for real-world problems, it frequently occurs that the PDF of the competing models is not known at all. The search for a sufficient statistic then becomes a moot point. Our results, however, indicate that a reasonable approach is to choose an estimator of the model parameters and then compare the PDFs of the sufficient statistics for the various models for the case of white Gaussian noise. This is termed the class-specific approach.

The paper is organized as follows. In Section II, the rationale for the decision rule is described. Section III applies the rule to the Gaussian linear model, whereas in Section IV, the class-specific form of the estimator is introduced. Specific signal processing examples for the linear model are discussed in Section V, whereas in Section VI, the results of a computer simulation are described. Finally, conclusions are given in Section VII.

II. Rationale for Decision Rule

It is well known that to minimize the probability of error P_e for a multiple hypothesis test one should use the maximum *a posterior* (MAP) rule [13]. In the model order selection problem, it is reasonable to assume no knowledge of the competing models. It is, therefore, usually assumed that the prior probabilities are equal. We will henceforth adopt this assumption. For M possible hypotheses with equal prior probabilities of $\Pr\{\mathcal{H}_i\} = 1/M$ of occurrence, the MAP rule reduces to the maximum likelihood (ML) rule. The latter chooses the hypothesis $\mathcal{H}_k$ to be the true hypothesis among $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_M\}$ if

$$p(\mathbf{x}|\mathcal{H}_k) > p(\mathbf{x}|\mathcal{H}_i) \qquad i \neq k$$

where $p(\mathbf{x}|\mathcal{H}_i)$ is the PDF of the data $\mathbf{x} = [x[0]\ x[1]\ \cdots\ x[N-1]]^T$ conditioned on $\mathcal{H}_i$ being true. The ML rule minimizes P_e or, equivalently, maximizes the probability of a correct decision $P_c = 1 - P_e$, which is

$$\begin{aligned} P_c &= \sum_{i=1}^{M} \Pr\{\text{decide } \mathcal{H}_i|\mathcal{H}_i\} \Pr\{\mathcal{H}_i\} \\ &= \frac{1}{M} \sum_{i=1}^{M} \int_{R_i} p(\mathbf{x}|\mathcal{H}_i) d\mathbf{x} \end{aligned} \tag{1}$$

where R_i is the decision region or subset of R^N for which we decide $\mathcal{H}_i$ if $\mathbf{x} \in R_i$. Note that the decision regions partition the space R^N. The ML rule assigns $\mathbf{x}$ to R_k if $p(\mathbf{x}|\mathcal{H}_k)$ is maximum. In practice, a major difficulty that arises in implementing this rule is that $p(\mathbf{x}|\mathcal{H}_i)$ is not known. It depends on the exact parameters for each assumed model.

To motivate the use of the CME, first, let the PDF of $\mathbf{x}$ depend on the model parameters $\boldsymbol{\theta}_i = [\theta_{i_1} \theta_{i_2} \cdots \theta_{i_{n_i}}]^T$, where $\boldsymbol{\theta}_i$ is $n_i \times 1$. Then, the PDF of the data can be written as $p(\mathbf{x};\ \boldsymbol{\theta}_i|\mathcal{H}_i)$. From (1), we seek to maximize

$$P_c = \frac{1}{M} \sum_{i=1}^{M} \int_{R_i} p(\mathbf{x};\ \boldsymbol{\theta}_i|\mathcal{H}_i)\, d\mathbf{x} \tag{2}$$

by choosing the decision regions R_i. The Bayesian approach would assign a prior PDF to $\boldsymbol{\theta}_i$ and then integrate it out. As discussed previously, this can lead to performance that is highly dependent on the priors assumed. We prefer to retain $\boldsymbol{\theta}_i$ as a deterministic but unknown parameter, i.e., the classical assumption, and *estimate* $\int_{R_i} p(\mathbf{x};\ \boldsymbol{\theta}_i|\mathcal{H}_i)\, d\mathbf{x}$ for each i. Then, we assign $\mathbf{x}$ to R_i if the estimated value is maximum. *Note that this approach more closely ties the decision procedure to its performance*. We next show that an optimal estimator of this quantity exists when the PDF family admits a complete sufficient statistic for the unknown parameters under each hypothesis. This optimal estimator is the MVU estimator.

Now, let the probability of a correct decision conditioned on $\mathcal{H}_i$ being true be denoted as

$$P_i(\boldsymbol{\theta}_i) = \int_{R_i} p(\mathbf{x};\ \boldsymbol{\theta}_i|\mathcal{H}_i)\, d\mathbf{x} \tag{3}$$

and using the indicator function defined as

$$I_i(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in R_i \\ 0, & \text{otherwise} \end{cases}$$

we have from (2) and (3) that

$$\begin{aligned} P_c &= \frac{1}{M} \sum_{i=1}^{M} P_i(\boldsymbol{\theta}_i) \\ &= \frac{1}{M} \sum_{i=1}^{M} \int I_i(\mathbf{x}) p(\mathbf{x};\ \boldsymbol{\theta}_i|\mathcal{H}_i)\, d\mathbf{x} \\ &= \frac{1}{M} \sum_{i=1}^{M} E(I_i(\mathbf{x})|\mathcal{H}_i). \end{aligned} \tag{4}$$

Next, consider the estimation of the expected value $E(I_i(\mathbf{x})|\mathcal{H}_i)$ that depends on $\boldsymbol{\theta}_i$. For ease of discussion, we simplify the notation to $E_X(I(\mathbf{x}))$, where X has the PDF $p_X(\mathbf{x};\ \boldsymbol{\theta})$, and $I(\mathbf{x})$ is the indicator function. Since the PDF of $\mathbf{x}$ depends on $\boldsymbol{\theta}$, we can let

$$P(\boldsymbol{\theta}) = E_X(I(\mathbf{x})) \tag{5}$$

and consider the estimation of $P(\boldsymbol{\theta})$, which is the probability that $\Pr\{\mathbf{x} \in R\}$ under the correct hypothesis. If a complete sufficient statistic $\mathbf{T}(\mathbf{x})$ exists for $\boldsymbol{\theta}$ and an unbiased estimator exists for $P(\boldsymbol{\theta})$, then the Rao–Blackwell–Lehmann–Scheffe theorem will yield the MVU estimator [12], [15]. Clearly, $I(\mathbf{x})$ is

an unbiased estimator since its expected value yields $P(\boldsymbol{\theta})$, as per (5). Thus, our MVU estimator for $P(\boldsymbol{\theta})$ is

$$\hat{P}(\boldsymbol{\theta}) = E_{X|T}(I(\mathbf{x})|\mathbf{T}(\mathbf{x}) = \mathbf{t}) \tag{6}$$

where the expectation is a conditional one with respect to the PDF of $\mathbf{x}$ once the sufficient statistic $\mathbf{T}(\mathbf{x})$ is observed. Note from the theory of sufficient statistics that $\hat{P}(\boldsymbol{\theta})$ will not depend on $\boldsymbol{\theta}$ but only on the data through $\mathbf{T}(\mathbf{x})$. Hence, $\hat{P}(\boldsymbol{\theta})$ depends only on $\mathbf{t}$. Now, using (6) in (4) and reintroducing the dependence on the assumed hypothesis, we have the estimate of P_c

$$\begin{aligned}\hat{P}_c &= \frac{1}{M}\sum_{i=1}^{M} E_{X|T_i}(I_i(\mathbf{x})|\mathbf{T}_i(\mathbf{x}) = \mathbf{t}_i) \\ &= \frac{1}{M}\sum_{i=1}^{M}\int I_i(\mathbf{x}) p_{X|T_i}(\mathbf{x}|\mathbf{T}_i(\mathbf{x}) = \mathbf{t}_i)d\mathbf{x}\end{aligned}$$

where $p_{X|T_i}$ is the conditional PDF of $\mathbf{x}$ conditioned on having observed $\mathbf{T}_i$ and which does not depend on $\boldsymbol{\theta}_i$. It should be noted that $\hat{P}_i(\boldsymbol{\theta}_i)$ is the MVU estimator of $P_i(\boldsymbol{\theta}_i)$ *only when* $\mathcal{H}_i$ *is true* and, thus, only when $\mathbf{T}_i$ is the *true* sufficient statistic. Otherwise, there is a misspecification [18].

We now propose the CME rule. This rule maximizes the *estimated* $\hat{P}_c$ by letting $\mathbf{x} \in R_i$ or, equivalently, $I_i(\mathbf{x}) = 1$ if $p_{X|T_i}$ is maximized. Hence, the CME rule decides $\mathcal{H}_k$ if

$$p_{X|T_k}(\mathbf{x}|\mathbf{T}_k(\mathbf{x}) = \mathbf{t}_k) > p_{X|T_i}(\mathbf{x}|\mathbf{T}_i(\mathbf{x}) = \mathbf{t}_i) \qquad i \neq k.$$

To actually compute the required conditional PDFs, we note that

$$p_{X|T_i}(\mathbf{x}|\mathbf{T}_i(\mathbf{x}) = \mathbf{t}_i) = \begin{cases} \dfrac{p_X(\mathbf{x};\, \boldsymbol{\theta}_i|\mathcal{H}_i)}{p_{T_i}(\mathbf{t}_i;\, \boldsymbol{\theta}_i|\mathcal{H}_i)}, & \text{if } \mathbf{T}_i(\mathbf{x}) = \mathbf{t}_i \\ 0, & \text{otherwise} \end{cases}$$

and therefore, the CME rule chooses the hypothesis that maximizes

$$L_{X|T_i}(\mathbf{x}) = \frac{p_X(\mathbf{x};\, \boldsymbol{\theta}_i|\mathcal{H}_i)}{p_{T_i}(\mathbf{T}_i(\mathbf{x});\, \boldsymbol{\theta}_i|\mathcal{H}_i)}. \tag{7}$$

Note, once again, that this conditional PDF does not depend on the unknown parameter vector $\boldsymbol{\theta}_i$ (the apparent dependence "cancels out") and therefore allows selection of a model without knowledge of the parameters of each model. It is interesting to observe that this rule is just the usual ML rule, except for the denominator term. In effect, the conditioning operation serves to discard the data within $\mathbf{x}$ that depends directly on the unknown parameter vector $\boldsymbol{\theta}_i$. This is because the data can be viewed as having been generated in two steps. In the first step, the sufficient statistic is generated by choosing a realization from a PDF that depends on the model parameters. In the second step, the remainder of the data is generated using a PDF that does not depend on the model parameters but only on the model dimension [4]. In essence, this is the fundamental concept of sufficiency. It is only the data obtained in step two that is used to make a decision. In Section III, an example illustrates this property.

In some ways, the CME is an extension of the concept of *similar regions* or regions of Neyman structure to multiple composite hypothesis testing [15]. Another connection is with the *conditionally principle* proposed by Fisher, which states that for statistical inference, only the part of the PDF that does not depend on any nuisance parameters should be used [15].

III. Application to the Linear Model

A. Unknown Noise Variance

An important application of the proposed model estimator is in choosing the order of the Gaussian linear model. Some specific examples of the Gaussian linear model are given in Section V along with a computer simulation. The Gaussian linear model is defined as [12]

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{8}$$

where

- $\mathbf{H}$ known $N \times p$ matrix;
- $\boldsymbol{\theta}$ $p \times 1$ parameter vector;
- $\mathbf{w}$ noise vector that is assumed to be Gaussian with mean zero and covariance matrix $\sigma^2\mathbf{I}$ or $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\, \sigma^2\mathbf{I})$.

The noise variance σ^2 is also assumed to be unknown. Hence, the unknown set of parameters is $[\boldsymbol{\theta}\sigma^2]^T$. In keeping with standard notation, we have not denoted this set of parameters as $\boldsymbol{\theta}$ since this is usually reserved for only the "signal" portion ($\mathbf{H}\boldsymbol{\theta}$) of the model. Hopefully, the meaning will become clear from the context. It is well known that this model for the data admits a sufficient statistic that is complete [10]. The sufficient statistic for the unknown parameters when properly normalized to make it an unbiased estimator is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \\ \hat{\sigma^2} &= \frac{1}{N-p}\left(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\right).\end{aligned}$$

It can further be shown that $\boldsymbol{\theta}$ and $\hat{\sigma}^2$ are independent random variables with corresponding PDFs

$$\begin{aligned}\hat{\boldsymbol{\theta}} &\sim \mathcal{N}(\boldsymbol{\theta},\, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) \\ u &= \frac{(N-p)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{N-p}.\end{aligned} \tag{9}$$

Hence, the sufficient statistic

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\sigma^2} \end{bmatrix} \tag{10}$$

can be used to implement the CME. To do so, we must determine from (7)

$$L_{X|T}(\mathbf{x}) = \frac{p_X(\mathbf{x};\, \boldsymbol{\theta},\, \sigma^2)}{p_T(\mathbf{T}(\mathbf{x});\, \boldsymbol{\theta},\, \sigma^2)}$$

where we have now omitted the i dependence for simplicity. This ratio is independent of the unknown parameters, as we now show in our computations. Using (8), we have

$$\begin{aligned}p_X(\mathbf{x};\, \boldsymbol{\theta},\, \sigma^2) = &\frac{1}{(2\pi\sigma^2)^{N/2}} \\ &\cdot \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})\right]\end{aligned}$$

and using the identity

$$(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}) = \left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)+\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T\mathbf{H}^T\mathbf{H}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \quad (11)$$

the PDF becomes

$$p_X(\mathbf{x};\,\boldsymbol{\theta},\,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)^T(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}})\right] \cdot\exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T\frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right]. \quad (12)$$

Now, for the sufficient statistic, we have from (9) and (10)

$$p_T(\mathbf{T};\,\boldsymbol{\theta},\,\sigma^2) = p_{\hat{\theta}}\left(\hat{\boldsymbol{\theta}};\,\boldsymbol{\theta},\,\sigma^2\right)p_{\hat{\sigma^2}}(\hat{\sigma^2};\,\sigma^2)$$

due to the independence, but from (9)

$$p_{\hat{\theta}}\left(\hat{\boldsymbol{\theta}};\,\boldsymbol{\theta},\,\sigma^2\right) = \frac{1}{(2\pi)^{p/2}\left|\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}\right|^{1/2}} \cdot\exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T\frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right] \quad (13)$$

and

$$p_U(u;\,\sigma^2) = \begin{cases} \dfrac{1}{2^{(N-p)/2}\,\Gamma\left(\dfrac{N-p}{2}\right)}\,u^{((N-p)/2)-1} \\ \quad\cdot\exp(-u/2), & u>0 \\ 0, & u<0 \end{cases}$$

so that

$$\hat{\sigma^2} = \frac{\sigma^2}{N-p}\,u$$

yields

$$p_{\hat{\sigma^2}}(\hat{\sigma^2};\,\sigma^2) = \frac{(N-p)/\sigma^2}{2^{(N-p)/2}\Gamma\left(\dfrac{N-p}{2}\right)}\left(\frac{(N-p)\hat{\sigma^2}}{\sigma^2}\right)^{((N-p)/2)-1} \cdot\exp\left(-\frac{1}{2}\,\frac{N-p}{\sigma^2}\,\hat{\sigma^2}\right). \quad (14)$$

Now, noting that

$$\hat{\sigma^2} = \frac{1}{N-p}\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)$$

we have from (12)–(14), after some simplifications

$$L_{X|T}(\mathbf{x}) = \frac{\Gamma\left(\dfrac{N-p}{2}\right)}{[\pi((N-p)]^{(N-p)/2}|\mathbf{H}^T\mathbf{H}|^{1/2}(\hat{\sigma^2})^{((N-p)/2)-1}}.$$

Upon taking natural logarithms and reinstating the i notation, the CME rule results. It says to choose the model order p that *minimizes*

$$\mathrm{CME}(i) = \frac{N-i-2}{2}\ln\hat{\sigma_i^2} + \left[\frac{1}{2}\ln|\mathbf{H}_i^T\mathbf{H}_i| + \ln\frac{[\pi(N-i)]^{(N-i)/2}}{\Gamma\left(\dfrac{N-i}{2}\right)}\right] \quad (15)$$

over $i = 1, 2, \ldots, M$, where M is the maximum possible order, and where

$$\hat{\sigma_i^2} = \frac{1}{N-i}\left(\mathbf{x}-\mathbf{H}_i\hat{\boldsymbol{\theta}}_i\right)\left(\mathbf{x}-\mathbf{H}_i\hat{\boldsymbol{\theta}}_i\right) = \frac{1}{N-i}\,\mathbf{x}^T(\mathbf{I}-\mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T)\mathbf{x}. \quad (16)$$

As claimed, the ratio is independent of the unknown parameters. It is seen that the first term is a fitting error, whereas the second term is a penalty for overfitting. Under fairly mild conditions, the CME, as given by (15), can be shown to be consistent. In Section V, we implement this model order estimator for some signal processing examples.

B. Known Noise Variance

We now determine the CME for the case of a linear model with *known* noise variance. Although it is of less practical importance than the previous case, we examine it to illustrate the essence of the CME. As alluded to earlier, the CME bases its decision on the part of the data that is not influenced by the model parameters. To see this, we have from (7) that

$$L_{X|T}(\mathbf{x}) = \frac{p_X(\mathbf{x};\,\boldsymbol{\theta})}{p_T(\mathbf{T}(\mathbf{x});\,\boldsymbol{\theta})}$$

where $\boldsymbol{\theta}$ refers to the signal parameters in the model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$. Since

$$\mathbf{T}(\mathbf{x}) = \hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta},\,\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$$

we have, upon using (11), the equation shown at the bottom of the next page, or

$$\begin{aligned} L_{X|T}(\mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}|\mathbf{H}^T\mathbf{H}|^{1/2}} \cdot\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)\right] \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \cdot\exp\left[-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}-\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T)\mathbf{x}\right] \cdot\frac{1}{|\mathbf{H}^T\mathbf{H}|^{1/2}} \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}}\exp\left[-\frac{1}{2\sigma^2}\,\mathbf{v}^T(\mathbf{x})\mathbf{v}(\mathbf{x})\right] \cdot\frac{1}{|\mathbf{H}^T\mathbf{H}|^{1/2}} \end{aligned} \quad (17)$$

where $\mathbf{v}(\mathbf{x}) = (\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T)\mathbf{x}$. The CME is obtained, therefore, by minimizing

$$\text{CME}(i) = \frac{\mathbf{v}_i^T(\mathbf{x})\mathbf{v}_i(\mathbf{x})}{2\sigma^2} + \frac{1}{2}\ln\left|\frac{\mathbf{H}_i^T\mathbf{H}_i}{2\pi\sigma^2}\right|$$

where $\mathbf{v}_i = (\mathbf{I} - \mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T)\mathbf{x}$ or equivalently

$$\text{CME}(i) = \frac{N-i}{2}\,\frac{\hat{\sigma_i^2}}{\sigma^2} + \frac{1}{2}\ln\left|\frac{\mathbf{H}_i^T\mathbf{H}_i}{2\pi\sigma^2}\right| \tag{18}$$

for $\hat{\sigma_i^2}$ given by (16).

Now, to relate $L_{X|T}(\mathbf{x})$ to the PDF of the data not dependent on $\boldsymbol{\theta}$, consider the following transformation from $\mathbf{x}$ to $[\mathbf{T}^T\mathbf{U}^T]^T$, where $\mathbf{T}$ is $p \times 1$ and $\mathbf{U}$ is $(N-p)\times 1$:

$$\begin{bmatrix}\mathbf{T}\\ \mathbf{U}\end{bmatrix} = \begin{bmatrix}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\\ \mathbf{B}\mathbf{x}\end{bmatrix} = \underbrace{\begin{bmatrix}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\\ \mathbf{B}\end{bmatrix}}_{\mathbf{A}}\mathbf{x} \tag{19}$$

and $\mathbf{B}$ is $(N-p)\times N$ and $\mathbf{BH} = \mathbf{0}$. The $N-p$ rows of $\mathbf{B}$ are chosen to span the orthogonal complement subspace of that spanned by the columns of $\mathbf{H}$. We furthermore assume that the basis of the former space is orthonormal so that $\mathbf{BB}^T = \mathbf{I}$. This decomposition is always possible. Now, because of the property that $\mathbf{BH} = \mathbf{0}$, the random vectors $\mathbf{T}$ and $\mathbf{U}$ are not only jointly Gaussian but independent as well. The latter follows by showing that the cross-covariance matrix is zero. Next, the Jacobian of the transformation of (19) is found. Again, due to $\mathbf{BH} = \mathbf{0}$, we have that

$$\mathbf{AA}^T = \begin{bmatrix}(\mathbf{H}^T\mathbf{H})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{BB}^T\end{bmatrix}$$

so that

$$|\mathbf{AA}^T| = |(\mathbf{H}^T\mathbf{H})^{-1}|\,|\mathbf{BB}^T| = \frac{1}{|\mathbf{H}^T\mathbf{H}|}$$

and since $|\mathbf{AA}^T| = |\mathbf{A}|^2$, we have that the Jacobian is

$$\left|\frac{\partial(\mathbf{t},\,\mathbf{u})}{\partial\mathbf{x}}\right| = |\mathbf{A}| = \frac{1}{|\mathbf{H}^T\mathbf{H}|^{1/2}}.$$

Hence, the PDF of $\mathbf{x}$ can be found using the inverse transformation of (19) as

$$\begin{aligned} p_X(\mathbf{x}) &= p_{TU}(\mathbf{t}(\mathbf{x}),\,\mathbf{u}(\mathbf{x}))\left|\frac{\partial(\mathbf{t},\,\mathbf{u})}{\partial\mathbf{x}}\right|\\ &= p_T(\mathbf{t}(\mathbf{x}))p_U(\mathbf{u}(\mathbf{x}))\left|\frac{\partial(\mathbf{t},\,\mathbf{u})}{\partial\mathbf{x}}\right|\\ &= p_T(\mathbf{t}(\mathbf{x}))\cdot\frac{p_U(\mathbf{u}(\mathbf{x}))}{|\mathbf{H}^T\mathbf{H}|^{1/2}}. \end{aligned}$$

Thus, we have finally that

$$L_{X|T}(\mathbf{x}) = \frac{p_U(\mathbf{u}(\mathbf{x}))}{|\mathbf{H}^T\mathbf{H}|^{1/2}} \tag{20}$$

which is the PDF of the $\mathbf{u}(\mathbf{x})$ data when adjusted by the Jacobian for the transformation to a random variable in $\mathbf{x}$ space. In addition, by expliciting determining the PDF of $\mathbf{u}(\mathbf{x})$ and substituting into (20), the result of (17) can be obtained.

It is easy to verify that the PDF of $\mathbf{u}(\mathbf{x})$ cannot depend on $\boldsymbol{\theta}$ since

$$\mathbf{u}(\mathbf{x}) = \mathbf{Bx} = \mathbf{BH}\boldsymbol{\theta} + \mathbf{Bw} = \mathbf{Bw}.$$

The CME is seen to use the part of the transformed data that does not depend on the model parameters, i.e., $\mathbf{u}(\mathbf{x}) = \mathbf{Bw}$ while discarding the part of the data, i.e., the sufficient statistic $\mathbf{T}(\mathbf{x})$, that does. In essence, $\mathbf{u}(\mathbf{x})$ forms the ancillary statistic, and its PDF, when adjusted for its dimensionality as seen in (20), forms the basis for discrimination.

IV. Class-Specific CME

The CME chooses the order that maximizes

$$\frac{p_X(\mathbf{x};\,\boldsymbol{\theta}_i|\mathcal{H}_i)}{p_{T_i}(\mathbf{T}_i(\mathbf{x});\,\boldsymbol{\theta}_i|\mathcal{H}_i)}.$$

As shown for the linear model and as is true in general, this ratio is independent of the value of $\boldsymbol{\theta}_i$ as long as the same value is used in the numerator and denominator. In some cases, it is advantageous to choose a specific value of $\boldsymbol{\theta}_i$. Consider the linear model example of Section III-A. There, we saw that

$$L_{X|T_i}(\mathbf{x}) = \frac{p_X(\mathbf{x};\,\boldsymbol{\theta}_i,\,\sigma^2|\mathcal{H}_i)}{p_T(\mathbf{T}_i(\mathbf{x});\,\boldsymbol{\theta}_i,\,\sigma^2|\mathcal{H}_i)}.$$

If we choose $\boldsymbol{\theta}_i = \mathbf{0}$ and $\sigma^2 = 1$, then this becomes

$$L_{X|T_i}(\mathbf{x}) = \frac{p_X(\mathbf{x};\,\mathbf{0},\,1|\mathcal{H}_i)}{p_T(\mathbf{T}_i(\mathbf{x});\,\mathbf{0},\,1|\mathcal{H}_i)}.$$

However

$$p_X(\mathbf{x};\,\mathbf{0},\,1|\mathcal{H}_i) = \frac{1}{(2\pi)^{N/2}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right) \tag{21}$$

for all i. Hence, in maximizing $L_{X|T_i}(\mathbf{x})$, we can omit the numerator term and just choose the hypothesis that *minimizes*

$$p_{T_i}(\mathbf{T}_i(\mathbf{x});\,\mathbf{0},\,1|\mathcal{H}_i).$$

From (21), however, we see that this is just the PDF of $\mathbf{T}_i(\mathbf{x})$ when the data $\mathbf{x}$ consists of white Gaussian noise with variance one. We term this somewhat fictitious hypothesis the $\mathcal{H}_0$ hypothesis. Hence, *the CME chooses the hypothesis that minimizes*

$$p_{T_i}(\mathbf{T}_i(\mathbf{x})|\mathcal{H}_0). \tag{22}$$

$$L_{X|T}(\mathbf{x}) = \frac{\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)\right]\exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right]}{\frac{1}{(2\pi)^{p/2}|\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}\exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right]}$$

We now show that minimizing (22) is equivalent to the CME rule. It should be pointed out that this result is quite important in that the determination of the exact PDF of a sufficient statistic can be difficult in practice. However, the problem has now been reduced to finding the PDF of the sufficient statistic, assuming that the data consist of white Gaussian noise. This being the *independent and identically distributed* (IID) case, the determination of the CME is simplified considerably. An example is that of model order estimation for an autoregressive process. This problem is addressed in [14].

From (10), we have that

$$\mathbf{T}_i(\mathbf{x}) = \begin{bmatrix} \hat{\boldsymbol{\theta}}_i \\ \hat{\sigma_i^2} \end{bmatrix}$$

and due to independence and using (13) and (14)

$$\begin{aligned} &\ln p_{T_i}(\mathbf{T}_i(\mathbf{x})|\mathcal{H}_0) \\ &\quad = \ln p_{\hat{\boldsymbol{\theta}}_i}\left(\hat{\boldsymbol{\theta}}_i;\ \mathbf{0},\ 1\right) + \ln p_{\hat{\sigma_i^2}}(\hat{\sigma_i^2};\ 1) \\ &\quad = \ln \frac{1}{(2\pi)^{i/2}|(\mathbf{H}_i^T\mathbf{H}_i)^{-1}|^{1/2}} \exp\left(-\frac{1}{2}\hat{\boldsymbol{\theta}}_i^T\mathbf{H}_i^T\mathbf{H}_i\hat{\boldsymbol{\theta}}_i\right) \\ &\qquad + \ln \frac{N-i}{2^{(N-i)/2}\Gamma\left(\dfrac{N-i}{2}\right)}[(N-i)\hat{\sigma_i^2}]^{((N-i)/2)-1} \\ &\qquad \cdot \exp\left[-\frac{1}{2}(N-i)\hat{\sigma_i^2}\right] \end{aligned}$$

where

$$\begin{aligned} \hat{\boldsymbol{\theta}}_i &= (\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T\mathbf{x} \\ \hat{\sigma_i^2} &= \frac{1}{N-i}\,\mathbf{x}^T(\mathbf{I} - \mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T)\mathbf{x} \end{aligned}$$

which, upon simplification, yields

$$\begin{aligned} \ln p_{T_i}(\mathbf{T}_i(\mathbf{x})|\mathcal{H}_0) &= \frac{N-i-2}{2}\ln\hat{\sigma_i^2} + \frac{1}{2}\ln|\mathbf{H}_i^T\mathbf{H}_i| \\ &\quad + \ln\frac{[\pi(N-i)]^{(N-i)/2}}{\Gamma\left(\dfrac{N-i}{2}\right)} \\ &\quad + \ln\frac{\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})}{(2\pi)^{N/2}} \\ &= \text{CME}(i) + \ln\frac{\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})}{(2\pi)^{N/2}}. \end{aligned}$$

This is clearly equivalent since the additional term does not depend on i. This approach follows from the class-specific ideas of [3].

Finally, the interpretation of the CME in this instance is of interest. The CME appears to choose the hypothesis that is least likely to resemble white Gaussian noise. As the model order is increased, the models take on more nonzero parameters, and hence, the "signal" departs more from zero, i.e., the white noise case. Once the correct order is reached, additional increases in order do not increase the signal but add only noise-like contributions. Thus, in the overfitting case, the model becomes more like white Gaussian noise, and its probability increases.

V. Specific Examples

Although there are many linear model examples that are important in signal processing, we choose to focus on two problems of interest. They are the estimation of a periodic signal in white Gaussian noise and the fitting of data by a polynomial. In either case, the results of the previous section apply. For the polynomial fitting problem, we give some computer simulation results and comparisons with existing approaches in Section VI.

Example 1—Periodic Signal in White Gaussian Noise: This problem is encountered in the processing of voiced speech in noise as well as detection of periodic signals in noise. What makes this problem a multiple composite hypothesis testing problem is the lack of knowledge of the period, in addition to the signal samples within a period. Assume then that we wish to estimate the period of the signal as well as the signal samples within a basic period. If $s[0],\ s[1]\ldots,\ s[p-1]$ are the samples within the first period of length p, we observe

$$\begin{aligned} x[n] &= s[n] + w[n] \qquad n = 0,\ 1,\ \ldots,\ N-1 \\ &= \begin{cases} s[n] + w[n] & 0 \le n \le p-1 \\ s[n-p] + w[n] & p \le n \le 2p-1 \\ \cdots & \cdots \end{cases} \end{aligned}$$

where $N > p$. This can be represented by the linear model as $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where

$$\mathbf{H}\boldsymbol{\theta} = \begin{bmatrix} \mathbf{I}_p \\ \mathbf{I}_p \\ \vdots \\ \mathbf{I}_p \\ \mathbf{J}_r \end{bmatrix} \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[p-1] \end{bmatrix}$$

where

$\mathbf{H}$	$N \times p$ matrix;
$\mathbf{I}_p$	$p \times p$ identity matrix;
$\mathbf{J}_r$	diag$(1,\ 1,\ \ldots,\ 1,\ 0,\ 0,\ \ldots,\ 0)$ with the first r diagonal elements equal to 1 and the remaining $p-r$ elements equal to zero;
N	$pK + r$ for K an integer that is the maximum number of periods.

If the period is known, then the MVU estimator of $\boldsymbol{\theta} = \mathbf{s}$ is just $\hat{\mathbf{s}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ or

$$\hat{s}[n] = \frac{1}{L_n}\sum_{l=0}^{L_n-1} x[n+lp] \qquad n = 0,\ 1,\ \ldots,\ p-1 \quad (23)$$

where L_n is the largest integer less than or equal to $(N-1-n)/p+1$, and the estimate of σ^2 is

$$\hat{\sigma^2} = \frac{1}{N-p}\sum_{n=0}^{N-1}(x[n] - \hat{s}[n])^2. \quad (24)$$

To determine the period, which is the number of signal model parameters, we use the CME given in Section IV. Note that the signal estimate $\hat{s}[n]$ used in (24) is that given in (23) for the first period and replicated for the remaining periods. In addition, for

TABLE I
MODEL ORDERS CHOSEN BY CME AND MDL—CORRECT ORDER IS 3

$N = 30$ and $\sigma^2 = 100$

order	1	2	**3**	4	5	6	7	8	9	10
MDL	0	26	847	84	39	4	0	0	0	0
CME	0	113	887	0	0	0	0	0	0	0

TABLE II
MODEL ORDERS CHOSEN BY CME AND MDL—CORRECT ORDER IS 3

$N = 30$ and $\sigma^2 = 10$

order	1	2	**3**	4	5	6	7	8	9	10
MDL	0	0	877	83	28	12	0	0	0	0
CME	0	0	999	1	0	0	0	0	0	0

TABLE III
MODEL ORDERS CHOSEN BY CME AND MDL—CORRECT ORDER IS 3

$N = 100$ and $\sigma^2 = 100$

order	1	2	**3**	4	5	6	7	8	9	10
MDL	0	0	964	34	2	0	0	0	0	0
CME	0	0	1000	0	0	0	0	0	0	0

this problem, $|\mathbf{H}^T\mathbf{H}| = (K+1)^r K^{p-r}$. Then, from (15), we have that

$$\begin{aligned}\mathrm{CME}(i) = {} & \frac{N-i-2}{2}\ln\left[\frac{1}{N-i}\sum_{n=0}^{N-1}(x[n]-\hat{s}_i[n])^2\right] \\ & + \tfrac{1}{2}\ln\left[(K_i+1)^{r_i}K_i^{i-r_i}\right] \\ & + \ln\frac{[\pi(N-i)]^{(N-i)/2}}{\Gamma\left(\dfrac{N-i}{2}\right)}\end{aligned}$$

where $N = K_i i + r_i$, and $i = 1, 2, \ldots, M$. The consistency conditions are easily shown to be satisfied, and hence, this is a consistent estimator. Other approaches to this problem are contained in [8], [13], and [19]. ♢

Example 2—Polynomial Fitting: Now, consider the fitting of a polynomial of unknown order. The data is assumed to consist of a polynomial signal embedded in white Gaussian noise of unknown variance. The polynomial signal is

$$s[n] = \theta_0 + \theta_1 n + \cdots + \theta_{p-1}n^{p-1}.$$

This data model is again a linear model with

$$\mathbf{H}\boldsymbol{\theta} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1^{p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & N-1 & \cdots & (N-1)^{p-1} \end{bmatrix}\begin{bmatrix}\theta_0 \\ \theta_1 \\ \vdots \\ \theta_{p-1}\end{bmatrix}.$$

The CME of (15) applies directly and can be shown to produce a consistent estimator. ♢

VI. COMPUTER SIMULATION RESULTS

For polynomial fitting as described in the previous section, we compare the performance of the CME and the minimum description length (MDL). The latter, which is based on an information theoretic formulation, determines the model order by minimizing

$$\mathrm{MDL}(i) = \frac{N}{2}\ln\check{\sigma}_i^2 + \frac{i+1}{2}\ln N$$

where

$$\check{\sigma}_i^2 = \frac{1}{N}\left(\mathbf{x}-\mathbf{H}_i\hat{\boldsymbol{\theta}}_i\right)^T\left(\mathbf{x}-\mathbf{H}_i\hat{\boldsymbol{\theta}}_i\right).$$

For large N, the two decision rules differ principally by the use of $(1/2)\ln|\mathbf{H}_i^T\mathbf{H}_i|$ for the CME penalty factor versus $((i+1)/2)\ln N$ for the MDL penalty factor. This has been pointed out in an asymptotic analysis by [7].

The polynomial chosen for the computer simulation is

$$s[n] = 0 + 0.4n + 0.1n^2 \qquad n = 0, 1, \ldots, N-1$$

so that the true order is $p = 3$. A maximum order of $M = 10$ is assumed. For $N = 30$ data points and $\sigma^2 = 100$, the results are shown in Table I. It is observed that whereas the MDL appears to overestimate the order, the CME tends to underestimate it. The probability of a correct model order selection, however, is larger for the CME. For a smaller amount of noise or $\sigma^2 = 10$, the results are shown in Table II. Now, the CME is almost perfect, whereas the MDL still overestimates the true order by a probability greater than 0.1. For the a large amount of noise or $\sigma^2 = 100$ but a longer data record $N = 100$, the CME is perfect, whereas the MDL still displays errors, as shown in Table III. Numerous other simulation examples confirm the superiority of the CME over the MDL.

VII. DISCUSSION AND CONCLUSIONS

The CME approach to model order selection has been shown to be a viable method in the absence of prior knowledge about model parameters for each competing model. Its justification is an attempt to choose a decision rule that is directly tied to maximizing the probability of a correct decision. In particular, we have examined its application to the Gaussian linear model and provided some signal processing examples. This application is easily derived. More complicated applications, such as the estimation of the model order of an autoregressive process, are currently being investigated. Compared with the MDL, the CME performance for finite data records appears to be superior. In addition, it can be proven that under fairly mild conditions, the CME is a consistent estimator of the model order. Its only

limitation is that the model PDF family must admit a set of sufficient statistics. They need not be a minimal set or even one of a dimension equal to that of the unknown parameters. In this case, the CME given by (7) is still independent of $\boldsymbol{\theta}$, but its interpretation as an MVU estimator will not hold. Approximate sufficient statistics such as the MLE can be used to extend its utility. However, this extension requires asymptotic arguments, which we have chosen to delay until a future paper.

REFERENCES

[1] H. Akaike, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, pp. 203–217, 1970.

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

[3] P. M. Baggenstoss, "Class-specific feature sets in classification," *IEEE Trans. Signal Processing*, vol. 47, pp. 3428–3432, Dec. 1999.

[4] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.

[5] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 1994.

[6] D. R. Cox and D. V. Hinkley, *Problems and Solutions in Theoretical Statistics*. New York: Chapman & Hall, 1978, pp. 160–162.

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

[8] T. W. Eddy, "Maximum likelihood detection and estimation for harmonic sets," *J. Acoust. Soc. Amer.*, pp. 149–155, 1980.

[9] F. Gustafsson and J. Hjalmarsson, "Twenty-one ML estimators for model selection," *Automatica*, vol. 31, pp. 1377–1392, 1995.

[10] F. A. Graybill, *Theory and Application of the Linear Model*. Duxbury, MA: Duxbury, 1976.

[11] R. L. Kashyap, "Optimal choise of AR and MA parts in autoregressive moving average model," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-3, pp. 99–104, 1982.

[12] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[13] ——, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[14] S. Kay, Baggenstoss, and A. Nuttall, "Approximation of probability density functions," *IEEE Trans. Signal Processing*, to be published.

[15] Sir M. Kendall and A. Stuart, *The Advanced Theory of Statistics*. New York: Macmillan, 1977, vol. 2.

[16] J. Rissanen, "Modeling by shortest data description," *Automatica*, pp. 465–478, 1978.

[17] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, pp. 461–464, 1978.

[18] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, pp. 1–26, 1982.

[19] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 418–423, 1976.

Steven Kay (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.E. degree from Columbia University, New York, NY, in 1973, and the Ph.D. degree from the Georgia Institute of Technology (Georgia Tech), Atlanta, 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 Tech to study comunications theory and digital signal processing. From 1977 to 1980, he was with the Submarine Signal Division, Raytheon Corp., Portsmouth, RI, where he engaged in research on autoregressive spectral estimation and design of sonar systems. He is currently Professor of electrical engineering at the University of Rhode Island, Kingston, and a consultant to industry and the United States Navy. He has written numerous papers, many of which have been reprinted in the IEEE Press book *Modern Spectral Analysis II*. He is a contributor to several edited books on spectral estimation and is the author of *Modern Spectral Estimation: Theory and Application* (Englewood Cliffs, NJ: Prentice-Hall, 1993). He conducts research in mathematical statistics with applications to digital signal processing. This includes the theory of detection, estimation, time series, and spectral analysis with applications to radar, sonar, communications, image processing, speech processing, biomedical signal processing, vibration, and financial data analysis.

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