Exponentially Embedded Families—New Approaches to Model Order Estimation

STEVEN KAY, Fellow, IEEE
University of Rhode Island

The use of exponential embedding of two or more probability density functions (pdfs) is introduced. Termed the exponentially embedded family (EEF) of pdfs, its properties are first examined and then it is applied to the problem of model order estimation. The proposed estimator is compared with the minimum description length (MDL) and is found to be superior for cases of practical interest. Also, we point out there is a relationship between the embedded family model order estimator and the generalized likelihood ratio test (GLRT). The embedded family estimator appears to extend the GLRT to the case of multiple alternative hypotheses that have differing numbers of unknown parameters.

Manuscript received August 3, 2004; released for publication October 6, 2004.

IEEE Log No. T-AES/41/1/844827.

Refereeing of this contribution was handled by J. A. Tague.

This work was supported by the Office of Naval Research under Contracts N00014-00-10871 and N00014-02-M-0166.

Author's address: Dept. of Electrical and Computer Engineering, University of Rhode Island, Kelly Hall, 4 E. Alumni Ave., Kingston, RI 02881, E-mail: (kay@ele.uri.edu).

0018-9251/05/$17.00 © 2005 IEEE

I. INTRODUCTION

Designing a model order estimator is a problem of considerable practical importance. Since the problem is essentially one of composite hypothesis testing, for which no optimal solution exists, there is no overall agreement on its solution. One common approach employs a Bayesian philosophy which assumes a noninformative prior in an effort to "integrate out" the unknown model parameters. Then, the effect of the prior is ignored [14, 17]. Along these lines the minimum description length (MDL) has been proposed based on coding arguments [16]. It also may be derived using asymptotic Bayesian arguments [17]. Recently, a new approach termed the conditional model estimator (CME) has been proposed and has been found to work quite well for polynomial model order estimation [15]. Unfortunately, it can produce poor results for problems in which the determinant of the Fisher information matrix does not result in a stringent enough penalty factor. In an effort to provide a more unified approach to composite hypothesis testing, we investigate the use of the exponentially embedded probability density function (pdf) family.

The concept of embedding pdfs in a more general family has appeared previously in an attempt to test between different pdf families [2]. However, there does not seem to be any detailed investigation into its properties and applications. Our main contribution is to present the embedded family in a more systematic form in an attempt to allow problems of composite hypothesis testing to be derived that perform well in practice. We term this method exponentially embedded families (EEFs). It allows the user to embed two or more pdfs into a family of pdfs that are indexed by one or more parameters. This new embedded family has the form of an exponential family and so inherits many of the nice mathematical and optimality properties of that family. It may also be viewed from the standpoint of differential geometry in that the embedded family forms a manifold in log-pdf space. Inference problems then become parameter tests and many new results obtained by using this viewpoint may be leveraged [3]. Our focus here, however, is to present the exponentially embedded family, and to apply it to the model order estimation problem. However, we also point out that the concepts presented lead to an extension of the generalized likelihood ratio test (GLRT) for multiple alternative hypotheses. Future work will concentrate on this important result.

The paper is organized as follows. In Section II we present the EEF while Section III gives some examples. Section IV introduces the reduced EEF. General results are given for the important linear model in Section V and a general asymptotic form of the reduced embedded family is given in Section VI. Section VII discusses model order estimation and a computer simulation of the proposed method is described in Section VIII. Finally, Section IX offers some conclusions.

II. DEFINITION OF EEF

Assume that we have two distinct pdfs $p_1(\mathbf{x})$ and $p_0(\mathbf{x})$, where $\mathbf{x} = [x[0]x[1]\ldots x[N-1]]^T$. These pdfs will later model the data under the general model hypothesis $\mathcal{H}_1$ and under a reference hypothesis $\mathcal{H}_0$, respectively. We wish to embed these pdfs in a family of pdfs, that is to say, the given pdfs will comprise two elements in a much larger set of pdfs. The family of pdfs will be denoted by $p(\mathbf{x};\eta)$, where η is the embedding parameter and takes on values $0 \le \eta \le 1$. The given pdfs are assigned to the "endpoints" as $p_1(\mathbf{x}) = p(\mathbf{x};1)$ and $p_0(\mathbf{x}) = p(\mathbf{x};0)$. There are many possible embeddings. For example, a common one is the mixture embedding defined as

$$p(\mathbf{x};\eta) = \eta p_1(\mathbf{x}) + (1-\eta)p_0(\mathbf{x}). \tag{1}$$

The domain of $p(\mathbf{x};\eta)$ with respect to $\mathbf{x}$ is the domain of its embedded pdfs, which are assumed to be the same. Note that $p(\mathbf{x};\eta)$ integrates to one for all $0 \le \eta \le 1$. For several reasons, which we subsequently enumerate, it is more fruitful to use an exponential embedding or

$$p(\mathbf{x};\eta) = \frac{p_1^{\eta}(\mathbf{x})p_0^{1-\eta}(\mathbf{x})}{\int p_1^{\eta}(\mathbf{x})p_0^{1-\eta}(\mathbf{x})d\mathbf{x}}. \tag{2}$$

We term this the binary EEF. It is a family of pdfs whose parameter η takes on all values in $[0,1]$ such that

$$M_0(\eta) = \int p_1^{\eta}(\mathbf{x})p_0^{1-\eta}(\mathbf{x})d\mathbf{x} < \infty. \tag{3}$$

It is clear that $M_0(0) = M_0(1) = 1$. Also, it can be shown using Holder's inequality that $M_0(\eta)$ is a convex function of η [1] and so $0 < M_0(\eta) \le \max(M_0(0), M_0(1)) = 1$. Thus, the EEF will be defined over $0 \le \eta \le 1$. Furthermore, by letting

$$K_0(\eta) = \ln M_0(\eta) \tag{4}$$

$$T(\mathbf{x}) = \ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \tag{5}$$

(2) can be written as

$$p(\mathbf{x};\eta) = \exp[\eta T(\mathbf{x}) - K_0(\eta) + \ln p_0(\mathbf{x})] \tag{6}$$

which is recognized as a one-parameter exponential family with natural parameter η [1, 2, 9, 10]. Also, the statistic $T(\mathbf{x})$ is a minimal sufficient statistic for η. The function $K_0(\eta)$ is termed the cumulant generating function since $M_0(\eta)$ is the moment generating function of $T(\mathbf{x})$, when $\mathbf{x}$ has the pdf $p_0(\mathbf{x})$. This is because from (3)

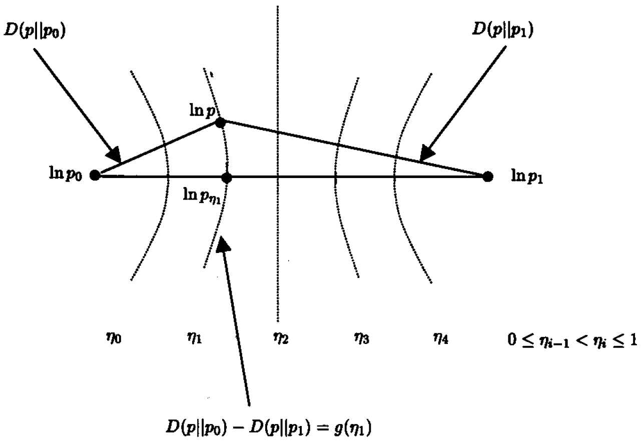

Fig. 1. Geometric interpretation of EEF as geodesic curve from $\ln p_0$ to $\ln p_1$.

$$\begin{aligned} M_0(\eta) &= \int \exp\left[\eta \ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\right] p_0(\mathbf{x})d\mathbf{x} \\ &= E_0[\exp(\eta T(\mathbf{x}))]. \end{aligned} \tag{7}$$

The subscript on the expectation operator denotes expectation with respect to $p_0(\mathbf{x})$.

We now discuss some motivations for our use of the EEF. First, if we examine its log-likelihood

$$\ln p(\mathbf{x};\eta) = \eta \ln p_1(\mathbf{x}) + (1-\eta)\ln p_0(\mathbf{x}) - K_0(\eta)$$

it is seen to be a convex combination of the given log-likelihoods. The factor $K_0(\eta)$ is of course needed for normalization. Thus, the log-likelihood is a smooth curve from $\ln p_0$ to $\ln p_1$ and may be thought of as comprising a submanifold of dimension one [3]. This curve may be interpreted as a geodesic in the log-likelihood space [4]. This interpretation depends on the geometric analogy between the Kullback-Liebler measure, or divergence, and squared Euclidean distance. The measure is defined as

$$D(p_1 \,\|\, p_0) = \int p_1(\mathbf{x}) \ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} d\mathbf{x}.$$

It is nonnegative and equals zero if and only if $p_0 = p_1$ for almost every $\mathbf{x}$. We refer to Fig. 1. There we envision a space of log-pdfs of which $\ln p_0$ and $\ln p_1$ are elements and which we seek to connect with a curve of minimum distance. To do so we first examine all log-pdfs along a single coordinate curve, say η_1. This curve is composed of all pdfs for which $D(p \,\|\, p_0) - D(p \,\|\, p_1) = g(\eta_1)$. The function g defines the locus of log-pdfs for which the difference of the distance-squares is constant and will depend on η. Since the divergence acts as a distance-squared, we interpret these curves as hyperbolas and thus are drawn as such. Now we choose the point on this curve as the one which minimizes the squared distance from $\ln p_0$ or which minimizes $D(p_{\eta_1} \,\|\, p_0)$. We then do this for all η to generate the geodesic from $\ln p_0$ to $\ln p_1$. Each point on the geodesic is therefore obtained by minimizing $D(p\,\|\,p_0)$ subject to the constraint that $D(p\,\|\,p_0) - D(p\,\|\,p_1) = g(\eta)$. The solution to this problem is well known and is given by (2) [5, 6] or equivalently the solution set is the line segment connecting $\ln p_0$ with $\ln p_1$. Second, $p(\mathbf{x};\eta)$ admits a sufficient statistic $T(\mathbf{x})$ for η that summarizes all the information for discrimination between p_1 and p_0. Hence, any decision procedures may be based on $p(\mathbf{x};\eta)$ without loss in performance [7]. Third, subject to regularity conditions, the exponential pdf is the only pdf to admit a sufficient statistic in the case of a single parameter and independent and identically distributed (IID) random variables [8]. Finally, as must be evident by now, the use of the EEF, which is an exponential family, enjoys a multitude of theoretical and practical properties as summarized in Appendix A. This makes the EEF amenable to a vast array of applications, one of which is explored in Section VII.

The many properties of the one-parameter exponential family extend to the multiparameter family. As such, we can define the M-ary EEF as

$$p(\mathbf{x};\eta) = \frac{p_1^{\eta_1}(\mathbf{x})p_2^{\eta_2}(\mathbf{x})\ldots p_{M-1}^{\eta_{M-1}}(\mathbf{x})p_0^{1-\sum_{i=1}^{M-1}\eta_i}(\mathbf{x})}{\int p_1^{\eta_1}(\mathbf{x})p_2^{\eta_2}(\mathbf{x})\ldots p_{M-1}^{\eta_{M-1}}(\mathbf{x})p_0^{1-\sum_{i=1}^{M-1}\eta_i}(\mathbf{x})d\mathbf{x}} \tag{8}$$

where $0 \le \eta_i \le 1$ and $\sum_{i=1}^{M-1}\eta_i \le 1$. We will not pursue this further here, but it will be the subject of a future paper.

III. EXAMPLES OF THE EEF

The first example of the EEF uses the linear model [11, 13] because of its wide applicability and its ease of exposition. The model is defined as

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

where $\mathbf{x}$ is an $N \times 1$ Gaussian random vector, $\mathbf{H}$ is a $N \times p$ constant matrix of full rank with $N > p$, and $\mathbf{w}$ is an $N \times 1$ Gaussian random vector with pdf $\mathbf{w} \sim \mathcal{N}(0,\sigma^2\mathbf{I})$. As such its pdfs are

$$p_1(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\right]$$
$$p_0(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x}\right]$$

where we associate the reference point, i.e., $\eta = 0$, with $\boldsymbol{\theta} = 0$. Thus,

$$\begin{aligned} T(\mathbf{x}) &= \ln\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}) + \frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x} \qquad (9) \\ &= \frac{\mathbf{x}^T\mathbf{H}\boldsymbol{\theta}}{\sigma^2} - \frac{1}{2}\frac{\|\mathbf{H}\boldsymbol{\theta}\|^2}{\sigma^2} \qquad (10) \end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm. Also,

$$\begin{aligned} K_0(\eta) &= \ln E_0[\exp(\eta T(\mathbf{x}))] \\ &= \ln[\exp(-(\eta/2)\|\mathbf{H}\boldsymbol{\theta}\|^2/\sigma^2)] \\ &\quad + \ln E_0[\exp((\eta\boldsymbol{\theta}^T\mathbf{H}^T/\sigma)(\mathbf{x}/\sigma))]. \end{aligned}$$

But for $\mathbf{u} \sim \mathcal{N}(0,\mathbf{I})$, $E(\exp(\boldsymbol{\xi}^T\mathbf{u})) = \exp(\boldsymbol{\xi}^T\boldsymbol{\xi}/2)$ and therefore

$$K_0(\eta) = \frac{\|\mathbf{H}\boldsymbol{\theta}\|^2}{2\sigma^2}(\eta^2 - \eta). \tag{11}$$

Substituting (10) and (11) into (6) and simplifying produces

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

Hence, if $p_0(\mathbf{x})$ is $\mathcal{N}(0,\sigma^2\mathbf{I})$, $p_1(\mathbf{x})$ is $\mathcal{N}(\mathbf{H}\boldsymbol{\theta},\sigma^2\mathbf{I})$, then $p(\mathbf{x};\eta)$ is $\mathcal{N}(\eta\mathbf{H}\boldsymbol{\theta},\sigma^2\mathbf{I})$. More generally, if p_1 is $\mathcal{N}(\boldsymbol{\mu}_1,\sigma^2\mathbf{I})$, p_0 is $\mathcal{N}(\boldsymbol{\mu}_0,\sigma^2\mathbf{I})$, then $p(\mathbf{x};\eta)$ is $\mathcal{N}(\eta\boldsymbol{\mu}_1 + (1-\eta)\boldsymbol{\mu}_0,\sigma^2\mathbf{I})$. The EEF is seen to consist of all $\mathcal{N}(\boldsymbol{\mu},\sigma^2\mathbf{I})$ pdfs whose mean lies on the line segment connecting $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$.

As a second example, consider p_1 as $\mathcal{N}(0,\mathbf{C}_1)$ and p_0 as $\mathcal{N}(0,\mathbf{C}_0)$. Then,

$$\begin{aligned} p(\mathbf{x};\eta) &= \frac{\left(\frac{1}{(2\pi)^{N/2}\det^{1/2}(\mathbf{C}_1)}\right)^{\eta}\exp\left[-\frac{1}{2}\eta\mathbf{x}^T\mathbf{C}_1^{-1}\mathbf{x}\right] \times \left(\frac{1}{(2\pi)^{N/2}\det^{1/2}(\mathbf{C}_0)}\right)^{1-\eta}\exp\left[-\frac{1}{2}(1-\eta)\mathbf{x}^T\mathbf{C}_0^{-1}\mathbf{x}\right]}{M_0(\eta)} \\ &= c\exp[-\tfrac{1}{2}\mathbf{x}^T(\eta\mathbf{C}_1^{-1} + (1-\eta)\mathbf{C}_0^{-1})\mathbf{x}] \end{aligned}$$

where c is a normalization constant. Hence, it is seen that $p(\mathbf{x};\eta)$ is $\mathcal{N}(0,\mathbf{C})$, where

$$\mathbf{C} = (\eta\mathbf{C}_1^{-1} + (1-\eta)\mathbf{C}_0^{-1})^{-1}.$$

The EEF consists of all $\mathcal{N}(0,\mathbf{C})$ pdfs whose information matrix $\mathbf{I} = \mathbf{C}^{-1}$ lies on a line segment connecting $\mathbf{I}_1 = \mathbf{C}_1^{-1}$ and $\mathbf{I}_0 = \mathbf{C}_0^{-1}$.

IV. REDUCED EEF FOR COMPOSITE HYPOTHESIS TESTING

Now the more interesting case arises when the signal pdf $p_1(\mathbf{x})$ contains some unknown parameters. In this case we replace $p_1(\mathbf{x})$ by $p_1(\mathbf{x};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown parameters. The resultant EEF will now depend on $\boldsymbol{\theta}$. Hence, we cannot test, for example, $\mathcal{H}_0 : p(\mathbf{x}) = p_0(\mathbf{x})$ versus $\mathcal{H}_1 : p(\mathbf{x}) \neq p_0(\mathbf{x})$ due to the unknown parameters $\boldsymbol{\theta}$ under $\mathcal{H}_1$. An alternative approach based on the EEF, which is now

$$p(\mathbf{x};\eta) = \exp\left[\eta\ln\left(\frac{p_1(\mathbf{x};\boldsymbol{\theta})}{p_0(\mathbf{x})}\right) - K_0(\eta) + \ln p_0(\mathbf{x})\right] \tag{12}$$

eliminates the dependence of the pdf on $\boldsymbol{\theta}$ by decomposing it into a term dependent on $\boldsymbol{\theta}$ and one that is independent of $\boldsymbol{\theta}$. This relies on the assumption that $p_1(\mathbf{x};\boldsymbol{\theta})$ admits a sufficient statistic for $\boldsymbol{\theta}$. This sufficient statistic need not be minimal and thus will always exist. A minimal sufficient statistic will sometimes exist but one can always use an approximate (actually asymptotically) minimal sufficient statistic which is given by the MLE of $\boldsymbol{\theta}$. Assuming a minimal sufficient statistic exists, we can use the Neyman-Fisher factorization theorem to write

$$p_1(\mathbf{x};\boldsymbol{\theta}) = p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}) \tag{13}$$

where $p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t})$ is the conditional pdf of $\mathbf{x}$, conditioned on the sufficient statistic $\mathbf{T}$, which is $p \times 1$ and $p_T(\mathbf{t};\boldsymbol{\theta})$ is the pdf of $\mathbf{T}$. The "conditional" pdf follows from (13) as

$$p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x})) = \frac{p_1(\mathbf{x};\boldsymbol{\theta})}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})}$$

and must be independent of $\boldsymbol{\theta}$. Note that the conditional pdf as defined is actually a density function over a manifold and so actual integration to determine probabilities will require the volume element of the manifold be included [18]. As a result we have from (12) and (13)

$$p(\mathbf{x};\eta) = \exp\left[\eta \ln\left(\frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p_0(\mathbf{x})}\right) + \eta \ln p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}) - K_0(\eta) + \ln p_0(\mathbf{x})\right].$$

Since $\boldsymbol{\theta}$ is completely unknown, inference cannot be made based on $p(\mathbf{x};\eta)$. In this type of situation, there are several approaches. Firstly, Fisher has argued that inference should be based on only the part of the likelihood function that is known or $p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))$, omitting the other term $p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})$ [12]. Note that if we do so, then $p(\mathbf{x};\eta)$ remains a valid pdf family over η (which is not so in the likelihood case), once $K_0(\eta)$ is adjusted for the correct normalization. In fact, it appears that the embedded family derives many of its useful properties from the use of a normalization factor. A second approach which ultimately leads to the same $p(\mathbf{x};\eta)$ is to assume that $p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})$ is a uniform pdf over its domain or to let $p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}) =$ constant. Although this is an improper pdf, the resulting $p(\mathbf{x};\eta)$ will still be a proper pdf. This approach is much the same as the use of improper pdfs in Bayesian analysis. However, it is important to note that this approach is not Bayesian in that $\boldsymbol{\theta}$ is a deterministic parameter. If we employ either of these arguments, the reduced EEF will be

$$p(\mathbf{x};\eta) = \exp\left[\eta \ln\left(\frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p_0(\mathbf{x})}\right) - K_0(\eta) + \ln p_0(\mathbf{x})\right] \tag{14}$$

where $K_0(\eta)$ is modified to be the new normalization factor. Since $K_0(\eta)$ has been modified, the domain of $p(\mathbf{x};\eta)$ may be a subset of the original one $0 \leq \eta \leq 1$. This is evident in the next example, which is included to demonstrate that the reduced EEF no longer depends on the unknown parameters but retains some discrimination necessary for hypothesis testing.

EXAMPLE 1 Reduced EEF for dc level in white Gaussian noise.

Consider the data $x[n] = A + w[n]$, where A is unknown and can take on values $-\infty < A < \infty$ and $w[n]$ is white Gaussian noise with known variance σ^2. Then,

$$p_1(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right]$$

and a minimal sufficient statistic for A is easily shown to be $T(\mathbf{x}) = (1/N)\sum_{n=0}^{N-1} x[n] = \bar{x}$ [13]. The conditional pdf required in (14) is

$$p_1(\mathbf{x} \mid T = t(\mathbf{x})) = \frac{p_1(\mathbf{x};A)}{p_T(t(\mathbf{x});A)}. \tag{15}$$

This ratio must be functionally independent of A. This is because if T is a sufficient statistic for A, then the pdf of the data $\mathbf{x}$ conditioned on the statistic cannot depend on the unknown parameter. To verify this we carry out the details. We first write the pdf as

$$p_1(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - \bar{x} + \bar{x} - A)^2\right]$$

and note that $\sum_{n=0}^{N-1}(x[n] - \bar{x} + \bar{x} - A)^2 = \sum_{n=0}^{N-1}(x[n] - \bar{x})^2 + N(\bar{x} - A)^2$ to yield

$$p_1(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}(x[n] - \bar{x})^2 + N(\bar{x} - A)^2\right)\right].$$

Also the pdf of $T = \bar{x}$ is $\mathcal{N}(A, \sigma^2/N)$ so that

$$\frac{p_1(\mathbf{x};A)}{p_T(t(\mathbf{x});A)} = \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}(x[n] - \bar{x})^2 + N(\bar{x} - A)^2\right)\right]}{\frac{1}{\sqrt{2\pi\sigma^2/N}} \exp\left[-\frac{1}{2\sigma^2/N}(\bar{x} - A)^2\right]}$$

$$= \frac{\sqrt{N}}{(2\pi\sigma^2)^{(N-1)/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - \bar{x})^2\right]$$

which is seen to be functionally independent of the unknown parameter A. Hence, the choice of A in (15) is arbitrary. By choosing $A = 0$ we associate $\eta = 0$

with $A = 0$, and thus

$$p_1(\mathbf{x} \mid T = t(\mathbf{x})) = \frac{p_1(\mathbf{x}; A = 0)}{p_T(t(\mathbf{x}); A = 0)} = \frac{p_0(\mathbf{x})}{p_T(t(\mathbf{x}); A = 0)}. \tag{16}$$

Hence we have that

$$\frac{p_1(\mathbf{x} \mid T = t(\mathbf{x}))}{p_0(\mathbf{x})} = \frac{1}{p_T(t(\mathbf{x}); A = 0)}.$$

Using this in (14) produces

$$p(\mathbf{x}; \eta) = \exp[-\eta \ln p_T(t(\mathbf{x}); A = 0) - K_0(\eta) + \ln p_0(\mathbf{x})] \tag{17}$$

which no longer depends on the unknown parameter A. To find $K_0(\eta)$ note that $T(\mathbf{x}) = \bar{x} \sim \mathcal{N}(0, \sigma^2/N)$ when $A = 0$ and thus

$$\ln p_T(t(\mathbf{x}); A = 0) = c - \frac{N\bar{x}^2}{2\sigma^2} \tag{18}$$

where c is a constant. We have

$$\begin{aligned} K_0(\eta) &= \ln E_0(\exp[-\eta \ln p_T(t(\mathbf{x}); A = 0)]) \\ &= -c\eta + \ln E_0[\exp((\eta/2)(N\bar{x}^2/\sigma^2))] \\ &= -c\eta + \ln \frac{1}{(1-\eta)^{1/2}} \end{aligned} \tag{19}$$

since $N\bar{x}^2/\sigma^2 \sim \chi_1^2$ for $p_0(\mathbf{x})$ the pdf of $\mathbf{x}$. Finally, we have from (17), (18), and (19)

$$p(\mathbf{x}; \eta) = \exp\left[(\eta/2)\frac{N\bar{x}^2}{\sigma^2} + \frac{1}{2}\ln(1-\eta) + \ln p_0(\mathbf{x})\right]. \tag{20}$$

As alluded to earlier, the domain of $p(\mathbf{x}; \eta)$ has been reduced to $0 \le \eta < 1$ since $\eta = 1$ is now excluded ($K_0(1) = \infty$). This reduced EEF pdf is recognized as a one-parameter exponential pdf. Also, by rearranging the terms it is easily shown that

$$p(\mathbf{x}; \eta) = \frac{\sqrt{1-\eta}}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I} - \eta\mathbf{P})\mathbf{x}\right]$$

where $\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ with $\mathbf{H} = [1\,1 \ldots 1]^T$. The matrix $\mathbf{P}$ is of course the $N \times N$ projection matrix for the linear model and has rank one. Furthermore, since

$$(\mathbf{I} - \eta\mathbf{P})^{-1} = \mathbf{I} + \frac{\eta}{1-\eta}\mathbf{P}$$

we see that the reduced EEF is $\mathcal{N}(0, \mathbf{C}_\eta)$ with

$$\mathbf{C}_\eta = \mathbf{I} + \frac{\eta}{1-\eta}\mathbf{P}.$$

To verify that the reduced EEF can be used for inference, we examine its form for $\eta = 0$ versus $\eta > 0$. If $\eta = 0$, the pdf corresponds to white Gaussian noise, while for $\eta > 0$, the pdf is that of N correlated Gaussian random variables. By a suitable linear transformation we can transform the pdf into one in which the random variables are Gaussian with zero means and are independent. The variances then become for $\eta = 0$, $\mathrm{var}(y[n]) = 1$ for $n = 1, 2, \ldots, N-1$ and for $\eta > 0$, $\mathrm{var}(y[n]) = 1$ for $n = 1, 2, \ldots, N-1$ and $\mathrm{var}(y[0]) = 1 + \eta/(1-\eta) = 1/(1-\eta)$. (Just note that $\mathbf{P}$ is symmetric and idempotent with one eigenvalue equal to one and the remaining eigenvalues equal to zero.) Hence, there is still discrimination available via the reduced pdf. However, it now manifests itself in the variance of the random variable $y[0]$ since inference is not possible based on the unknown mean A.

V. REDUCED EEF FOR THE LINEAR MODEL

To generalize Example 1 we determine the EEF for the linear model. This result will be used later. In the case of the linear model it is well known that the minimal sufficient statistic for $\boldsymbol{\theta}$, which is also the maximum likelihood estimator (MLE), is [13]

$$\mathbf{T}(\mathbf{x}) = \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

and has the pdf $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. Thus, from (14) and using the same arguments as before

$$\begin{aligned} T_\eta(\mathbf{x}) &= \ln \frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p_0(\mathbf{x})} \\ &= \ln \frac{p_1(\mathbf{x} \mid \hat{\boldsymbol{\theta}})}{p_0(\mathbf{x})} \\ &= \ln \frac{p_1(\mathbf{x}; \boldsymbol{\theta})}{p(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})p_0(\mathbf{x})}. \end{aligned}$$

Since the ratio is independent of $\boldsymbol{\theta}$, we choose $\boldsymbol{\theta} = 0$ so that $p_1(\mathbf{x}; \boldsymbol{\theta} = 0) = p_0(\mathbf{x})$ and we have

$$T_\eta(\mathbf{x}) = -\ln p(\hat{\boldsymbol{\theta}}; 0)$$

where $p(\hat{\boldsymbol{\theta}}; 0)$ is the pdf of $\hat{\boldsymbol{\theta}}$ obtained for the choice of $\boldsymbol{\theta} = 0$. Ignoring the constant terms (not dependent on $\mathbf{x}$) since these are absorbed by $K_0(\eta)$, this becomes

$$T_\eta(\mathbf{x}) = \frac{\hat{\boldsymbol{\theta}}^T\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}}{2\sigma^2} \tag{21}$$

and therefore

$$K_0(\eta) = \ln\left[E_0\left(\exp\left(\eta\frac{\hat{\boldsymbol{\theta}}^T\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}}{2\sigma^2}\right)\right)\right].$$

But $\hat{\boldsymbol{\theta}}^T\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}/\sigma^2 \sim \chi_p^2$ for $\boldsymbol{\theta} = 0$ and so

$$K_0(\eta) = \ln\left[\frac{1}{(1-\eta)^{p/2}}\right]. \tag{22}$$

From (14), (21), and (22) the EEF for the linear model becomes

$$p(\mathbf{x}; \eta) = \exp\left[\eta\frac{\hat{\boldsymbol{\theta}}^T\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}}{2\sigma^2} + \frac{p}{2}\ln(1-\eta) + \ln p_0(\mathbf{x})\right]. \tag{23}$$

VI. GENERAL REDUCED EEF VIA ASYMPTOTICS

In extending the utility of the results for the linear model to more general problems one can make use of the asymptotic properties of the MLE. Specifically, if the pdf $p_1(\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$ has unknown parameters $\boldsymbol{\theta}$, a $p \times 1$ vector, and satisfies certain regularity conditions, then the MLE $\hat{\boldsymbol{\theta}}$ is asymptotically a minimal sufficient statistic. It is shown in Appendix B that asymptotically we have the reduced EEF

$$p(\mathbf{x};\eta) = \exp[(\eta/2)l_G(\mathbf{x}) + (p/2)\ln(1-\eta) + \ln p(\mathbf{x};\boldsymbol{\theta}_0)] \tag{24}$$

where $0 \le \eta < 1$ and $l_G(\mathbf{x})$ is the GLRT statistic given by

$$l_G(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}})}{p(\mathbf{x};\boldsymbol{\theta}_0)}. \tag{25}$$

It can be shown that (24) is also valid in the presence of nuisance parameters $\boldsymbol{\alpha}$ if we use as the GLRT statistic

$$l_G(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\alpha}}_1)}{p(\mathbf{x};\boldsymbol{\theta}_0,\hat{\boldsymbol{\alpha}}_0)} \tag{26}$$

where $\hat{\boldsymbol{\alpha}}_1$ is the unconstrained MLE of $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}_0$ is the constrained (when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$) MLE of $\boldsymbol{\alpha}$.

VII. AN APPLICATION TO MODEL ORDER ESTIMATION

One approach to model order estimation is to embed all the models into an M-ary EEF as given by (8). Then, reducing it to a family dependent only on $\eta_1,\eta_2,\ldots,\eta_{M-1}$ by using the sufficient statistics, we can formulate a model order estimator. This will be reported on in a future paper. Here, we retain the binary embedding for which the EEF for the ith order model is from (6)

$$p_i(\mathbf{x};\eta) = \exp\left[\eta\ln\frac{p_i(\mathbf{x};\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} - K_0(\eta) + \ln p_0(\mathbf{x})\right]$$

for $i = 1,2,\ldots,M-1$ and $\boldsymbol{\theta}_i$ is an $i \times 1$ vector of unknown model parameters. As usual $p_0(\mathbf{x})$ is assumed known. Using the sufficient statistics to yield a reduced EEF and letting $\boldsymbol{\theta}_i = 0$ we have

$$p_i(\mathbf{x};\eta) = \exp\left[\eta\ln\left(\frac{1}{p_{T_i}(\mathbf{t}(\mathbf{x});0)}\right) - K_0(\eta) + \ln p_0(\mathbf{x})\right].$$

The proposed model order estimator chooses the model whose maximized likelihood function or $p_i(\mathbf{x};\hat{\eta})$ is maximum over all models. The estimator $\hat{\eta}$ is then the MLE for η based on the ith order model. Thus, we choose model k if

$$\mathrm{EEF}(i) = \max_{\eta}\left[\eta\ln\left(\frac{1}{p_{T_i}(\mathbf{t}(\mathbf{x});0)}\right) - K_0(\eta)\right] \tag{27}$$

is maximized for $i = k$. This approach is justified by arguing that for large data records, i.e., asymptotically, the divergence between the true pdf and the estimated one becomes a minimum. The argument is given in Appendix C.

As an illustration, we again consider the linear model. Then, using (23) and replacing p with i, we have

$$\begin{aligned}\eta\ln\left(\frac{1}{p_{T_i}(\mathbf{t}(\mathbf{x});0)}\right) - K_0(\eta) &= \eta\frac{\hat{\boldsymbol{\theta}}_i^T\mathbf{H}_i^T\mathbf{H}_i\hat{\boldsymbol{\theta}}_i}{2\sigma^2} + \frac{i}{2}\ln(1-\eta)\\ &= \eta\frac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{2\sigma^2} + \frac{i}{2}\ln(1-\eta)\end{aligned} \tag{28}$$

where $\mathbf{P}_{H_i} = \mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T$ and $\mathbf{H}_i$ is $N \times i$. Maximizing over the domain $0 \le \eta < 1$ by differentiating and setting equal to zero produces the global maximum (since $-K_0(\eta)$ is concave as summarized in Appendix A)

$$\hat{\eta} = \begin{cases} 1 - \dfrac{1}{\dfrac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{i\sigma^2}} & \dfrac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{i\sigma^2} > 1\\[2ex] 0 & \dfrac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{i\sigma^2} \le 1. \end{cases} \tag{29}$$

Since $\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x} \ge 0$, we must have $\hat{\eta} \le 1$. But if $\hat{\eta} = 0$, then EEF(i) = 0 and otherwise EEF(i) > 0 as we now show. For $\hat{\eta} > 0$ and letting $Q_i = \mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}/\sigma^2$ we have from (27), (28), and (29)

$$\mathrm{EEF}(i) = \left(1 - \frac{i}{Q_i}\right)\frac{Q_i}{2} + \frac{i}{2}\ln\frac{i}{Q_i} = \frac{Q_i}{2} - \frac{i}{2}\left(\ln\frac{Q_i}{i} + 1\right)$$

or equivalently letting EEF(i) be multiplied by two

$$\mathrm{EEF}(i) = Q_i - i\left(\ln\frac{Q_i}{i} + 1\right).$$

But for $\hat{\eta} > 0$, we have that $Q_i > i$ and thus EEF(i) is monotonically increasing in Q_i and equals zero for $Q_i = i$.

We can also write the EEF in more compact form as

$$\mathrm{EEF}(i) = \left(Q_i - i\left(\ln\frac{Q_i}{i} + 1\right)\right)u\left(\frac{Q_i}{i} - 1\right) \tag{30}$$

where $u(x)$ is the unit step function. Note that Q_i increases with model order while $i(\ln(Q_i/i) + 1)$, the penalty term, will also usually increase with model order. It is interesting to note that the usual determinant of the Fisher information matrix is not present in the penalty term. The EEF model order estimator will then exhibit an increase in robustness over such approaches as the asymptotic MAP [14], and CME [15], all of which have this term present. This is because the determinant of the Fisher information matrix is not guaranteed to increase and so is not in general a penalty factor. As an example, in the linear model part of the penalty term is $\ln\det(\mathbf{I}(\boldsymbol{\theta}_i)) = \ln\det(\mathbf{H}_i^T\mathbf{H}_i/\sigma^2)$ and it can be shown that

$$\ln\det\left(\frac{\mathbf{H}_i^T\mathbf{H}_i}{\sigma^2}\right) = \ln(\mathbf{h}_i^T\mathbf{P}_{H_{i-1}}^{\perp}\mathbf{h}_i) + \ln\det\left(\frac{\mathbf{H}_{i-1}^T\mathbf{H}_{i-1}}{\sigma^2}\right)$$

where $\mathbf{h}_i$ is the new column vector of $\mathbf{H}_i$ and $\mathbf{P}_{H_{i-1}}^{\perp}$ is the orthogonal projection operator. If the new column lies within the same subspace as that spanned by the columns of $\mathbf{H}_{i-1}$, then $\mathbf{h}_i^T\mathbf{P}_{H_{i-1}}^{\perp}\mathbf{h}_i = 0$ and the Fisher information matrix for the ith order model will be singular [13]. The penalty will then will be $-\infty$ and the estimator will always choose that model. This appears to be at odds with what one would normally expect.

The MDL [16] for the same problem can be shown to be

$$\mathrm{MDL}(i) = -Q_i + i\ln N. \tag{31}$$

The main difference is seen to be the penalty term. For the EEF approach it is

$$P(i) \approx i\ln\frac{Q_i}{i}$$

The two will be about the same if $E(Q_i)/i = O(N)$. However, for a polynomial fitting problem, as an example, we will have $E(Q_i)/i = O(N^{\alpha})$, where $\alpha > 1$, and so the EEF will have a more stingent penalty factor. We would therefore expect a lower probability of overparametrization. This is borne out in the simulation examples of Section VIII. The penalty term in the EEF approach depends on Q_i/i or the energy to noise ratio (ENR) per dimension. Finally, it can be shown that the EEF model order estimator is consistent.

To apply the model order estimator more generally we can use an asymptotic argument. As shown in Appendix D the model order estimator chooses the kth model if

$$\mathrm{EEF}(i) = \left(l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right]\right) u\left(\frac{l_{G_i}(\mathbf{x})}{i} - 1\right) \tag{32}$$

where

$$l_{G_i}(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}}_i)}{p(\mathbf{x};\boldsymbol{\theta}_0)}$$

is maximum for $i = k$. Here $\hat{\boldsymbol{\theta}}_i$ is the MLE for the parameters of the ith order model. It can furthermore be shown that in the presence of nuisance parameters $\boldsymbol{\alpha}$ we obtain (32) but with

$$l_{G_i}(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}}_i,\hat{\boldsymbol{\alpha}}_i)}{p(\mathbf{x};\boldsymbol{\theta}_0,\hat{\boldsymbol{\alpha}}_0)} \tag{33}$$

where $\hat{\boldsymbol{\alpha}}_i$ is the unconstrained MLE for the ith order model and $\hat{\boldsymbol{\alpha}}_0$ is the constrained MLE (for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$).

It is interesting to note that the EEF extends the GLRT to allow testing of multiple alternative hypotheses, and in particular, when the alternatives have differing numbers of unknown parameters. For example, if we wish to implement the usual binary test of $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\alpha}$ versus $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \boldsymbol{\alpha}$, then since

$$\mathrm{EEF}(i) = \left(l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right]\right) u\left(\frac{l_{G_i}(\mathbf{x})}{i} - 1\right)$$

with

$$l_{G_i}(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}}_i,\hat{\boldsymbol{\alpha}}_i)}{p(\mathbf{x};\boldsymbol{\theta}_0,\hat{\boldsymbol{\alpha}}_0)}$$

we need only compute the value of the EEF with $i = p$ and compare it to a threshold. But the EEF is a monotonic function of $l_{G_i}(\mathbf{x})$ and so an equivalent test is to decide $\mathcal{H}_1$ if $l_{G_i}(\mathbf{x}) > \gamma$, which is the usual GLRT. When there are multiple alternative hypotheses, however, the EEF first computes the GLRT for each hypothesis or $l_{G_i}(\mathbf{x})$, next applies the transformation

$$g_i(x) = \left(x - i\left[\ln\left(\frac{x}{i}\right) + 1\right]\right) u\left(\frac{x}{i} - 1\right)$$

and then chooses the hypothesis that yields the maximum. The effect of the transformations $g_i(x)$ is to penalize the $l_{G_i}(\mathbf{x})$ as the number of tested parameters increases. It is easily shown that $g_i(x)$ is a monotonic function and $g_j(x) < g_i(x)$ for $j > i$. To set a constant probability of false alarm we can next compare the EEF statistic (the one that yielded the maximum) to a threshold. Hence, it would appear that the EEF model order estimator as described herein extends the GLRT to the case of multiple alternative hypotheses. The usual problem of the GLRT, that the hypothesis with the most number of unknown parameters will always be chosen, has been solved by first applying the transformations $g_i(x)$.

As a second example of interest, we can apply the asymptotic form to the problem of model order estimation for an autoregressive (AR) process [19]. We use (32) and (33). In this case the unknown parameters $\boldsymbol{\theta}$ are the AR filter parameters and the nuisance parameter is the excitation noise variance σ_u^2. We therefore have that $\boldsymbol{\theta} = \mathbf{a} = [a[1]\,a[2]\ldots a[i]]^T$ and $\boldsymbol{\alpha}_i = \sigma_{u_i}^2$ for the ith order model. Hence, $\hat{\boldsymbol{\theta}}_i$ is the MLE of the AR filter parameters and $\hat{\boldsymbol{\alpha}}_i$ is the MLE of the excitation noise variance for the ith order model. For the reference value of $\boldsymbol{\theta}$ we take $\boldsymbol{\theta}_0 = 0$ or the AR filter parameters are all zero so that the reference hypothesis is just white Gaussian noise with unknown variance. The exact pdf for an AR process is difficult to work with directly so that as an approximation (valid for large N) we use the conditional pdf [19]

$$p(\mathbf{x};\mathbf{a},\sigma^2) = \frac{1}{(2\pi\sigma_u^2)^{(N-i)/2}}\exp\left[-\frac{1}{2\sigma_u^2}\sum_{n=i}^{N-1}\left(x[n] + \sum_{k=1}^{i}a[k]x[n-k]\right)^2\right] \tag{34}$$

and for the reference hypothesis $\mathbf{a} = 0$ so that

$$p(\mathbf{x};\mathbf{a}=0,\sigma^2) = \frac{1}{(2\pi\sigma_u^2)^{(N-i)/2}}\exp\left[-\frac{1}{2\sigma_u^2}\sum_{n=i}^{N-1}x^2[n]\right]. \tag{35}$$

The required $l_{G_i}(\mathbf{x})$ for these pdfs is shown in Appendix E to be

$$l_{G_i}(\mathbf{x}) = (N-i)\ln\left(\frac{\hat{\sigma}_{u_0}^2}{\hat{\sigma}_{u_i}^2}\right) \tag{36}$$

where

$$\hat{\sigma}_{u_i}^2 = \frac{1}{N-i}\mathbf{x}_i^T(\mathbf{I}-\mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T)\mathbf{x}_i$$

$$\hat{\sigma}_{u_0}^2 = \frac{1}{N-i}\mathbf{x}_i^T\mathbf{x}_i$$

and

$$\mathbf{x}_i = [x[i]x[i+1]\ldots x[N-1]]^T$$

$$\mathbf{H}_i = \begin{bmatrix} x[i-1] & x[i-2] & \ldots & x[0] \\ x[i] & x[i-1] & \ldots & x[1] \\ \vdots & \vdots & \vdots & \vdots \\ x[N-2] & x[N-3] & \ldots & x[N-1-i] \end{bmatrix}.$$

For this problem the MDL can be shown to be [19]

$$\mathrm{MDL}(i) = N\ln(\hat{\sigma}_{u_i}^2) + i\ln(N). \tag{37}$$

Computer simulation results are given in the next section.

VIII. COMPUTER SIMULATIONS

A. Polynomial Model Order Estimation

The problem of determining the order of a polynomial that is embedded in white Gaussian noise is a problem of much interest. Since this falls within the class of linear models, we can apply our results directly. Hence, we compare the MDL or (31) to the EEF or (30) and also the CME as described in [15]. We therefore choose the model order to maximize

$$-\mathrm{MDL}(i) = \frac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{\sigma^2} - i\ln N$$

$$\mathrm{EEF}(i) = \left(\frac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{\sigma^2} - i\left[\ln\left(\frac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{i\sigma^2}\right)+1\right]\right)$$
$$\times u\left(\frac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}/\sigma^2}{i} - 1\right)$$

$$-\mathrm{CME}(i) = \frac{\mathbf{x}^T\mathbf{P}_{H_i}\mathbf{x}}{\sigma^2} - \ln\det\left(\frac{\mathbf{H}_i^T\mathbf{H}_i}{2\pi\sigma^2}\right)$$

for $i = 1,2,\ldots,p_{\max}$.

Consider the parabolic signal

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

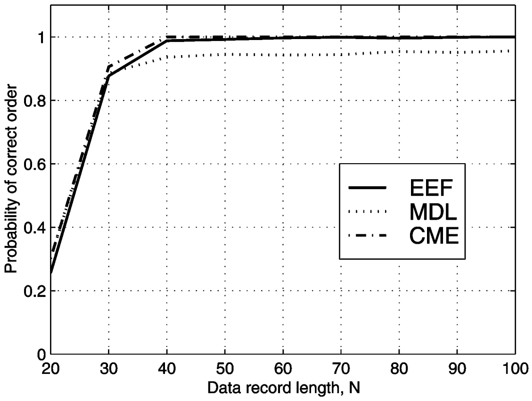

Fig. 2. Probability of correct model order for polynomial versus data record length.

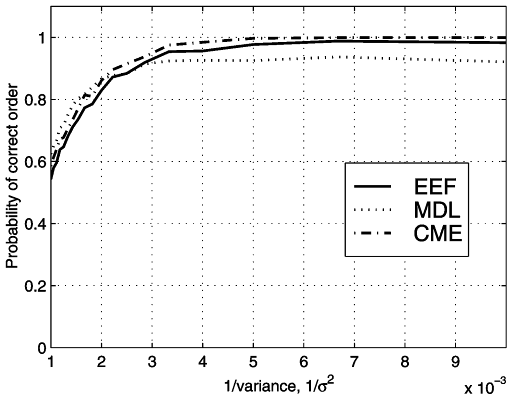

Fig. 3. Probability of correct model order for polynomial versus inverse noise variance.

For a noise variance of $\sigma^2 = 100$ the probability of correct model order or P_c is shown in Fig. 2 versus the data record length N. The CME appears to produce the best results with the EEF nearly as good. The MDL is consistently poorer owing to its inappropriate penalty factor for a polynomial. If the data record is kept fixed at $N = 40$, then P_c versus the inverse noise variance or $1/\sigma^2$ is shown in Fig. 3. Again the CME outperforms the EEF but only marginally while the MDL is consistently poorer.

B. CME Difficulties for Nonincreasing Determinant of Fisher Information Matrix

For a sinusoidal signal the determinant of the Fisher information matrix will not necessarily increase with model order or if it does, it may increase very slowly. Hence, as shown in the next simulation the CME may not produce good results. The example to follow is of limited practical interest but only serves to highlight the potential difficulties of the CME. The signal consists of three sinusoids embedded in white Gaussian noise or

$$x[n] = \cos(2\pi(0.1)n) + \cos(2\pi(0.11)n)$$
$$+\cos(2\pi(0.12)n) + w[n], \qquad n = 1,2,\ldots,N-1$$

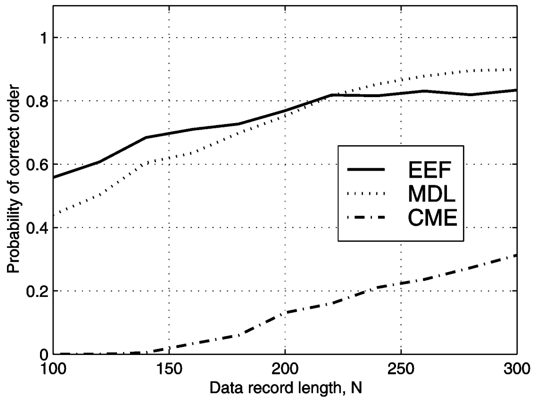

Fig. 4. Probability of correct model order for sinusoids versus data record length.

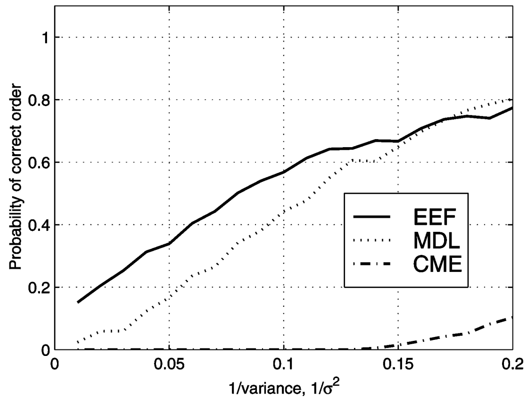

Fig. 5. Probability of correct model order for sinusoids versus inverse noise variance.

so that the true order or number of sinusoids is three. As before the signal model has the linear model form since the frequencies are assumed known. However, the amplitudes and phases in addition to the number of sinusoids is assumed unknown. We assume that the models are nested so that we test the models with sinusoidal frequencies $f_1 = 0.10$ for model 1, $f_1 = 0.10, f_2 = 0.11$ for model two, $f_1 = 0.10, f_2 = 0.11, f_3 = 0.12$ for model three,..., $f_1 = 0.10, \ldots, f_8 = 0.17$ for model eight. In Fig. 4 the P_c versus N is shown when $\sigma^2 = 10$. Now the CME performs very poorly. This is because the penalty factor is incorrect for short data records for this example. We have that $\mathbf{H}_i^T\mathbf{H}_i \approx (N/2)\mathbf{I}_i$ and therefore

$$\begin{aligned}\ln\det\left(\frac{\mathbf{H}_i^T\mathbf{H}_i}{2\pi\sigma^2}\right) &\approx \ln\det\left(\frac{N}{4\pi\sigma^2}\mathbf{I}_i\right)\\ &= \ln\left(\frac{N}{4\pi\sigma^2}\right)^i\\ &= i\ln\left(\frac{N}{4\pi\sigma^2}\right).\end{aligned}$$

This will only increase with i if $N/(4\pi\sigma^2) > 1$, which for this example requires $N > 40\pi \approx 125$. For smaller data records the penalty factor will actually decrease with i. This is evident in Fig. 4. The EEF outperforms the MDL for $N < 230$ but is slightly poorer for larger data records. The performance versus the inverse noise variance for $N = 100$ is shown in Fig. 5. Note again that for the penalty factor to be increasing we must have $N/(4\pi\sigma^2) > 1$ or $1/\sigma^2 > (4\pi)/N \approx 0.126$. Again the EEF outperforms the MDL for smaller inverse noise variance (equivalently for lower energy-to-noise ratio). It is only when the energy-to-noise ratio is very large or $NA^2/(2\sigma^2) = 100/(2(0.2)) = 24$ dB that the performance of the MDL is slightly better.

C. AR Model Order Estimation

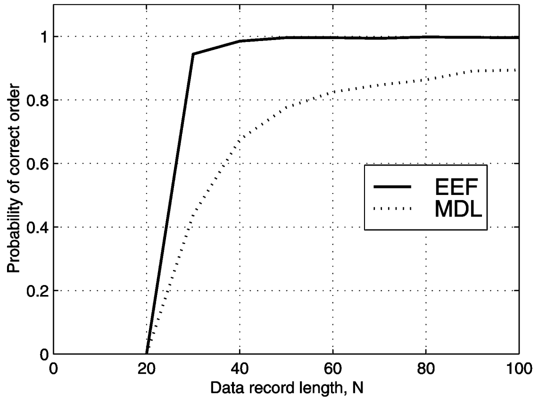

Fig. 6. Probability of correct model order for AR process of order 4.

Since the CME is in general unreliable, we only compare the EEF to the MDL for the problem of AR model order estimation. An AR process of order 4 whose parameters are given by

$$\begin{aligned}a[1] &= -2.760\\ a[2] &= 3.809\\ a[3] &= -2.654\\ a[4] &= 0.924\\ \sigma_u^2 &= 1\end{aligned}$$

is used to determine the performance of the EEF as given by (32) and (36), and for the MDL estimator given by (37). The results are shown in Fig. 6 versus the data record length. As is evident the EEF outperforms the MDL for all data record lengths.

It appears that overall the EEF produces good model order estimation, especially for short data records and/or low energy-to noise-ratios. This is the regime when the performance of a model order estimator is most critical.

IX. CONCLUSIONS

A new approach to composite hypothesis testing has been proposed. It embeds two different pdfs into a new pdf family that is an exponential family. In particular, an application to model order estimation has been examined in detail and the proposed approach based on the EEF is shown to perform quite well. Also, an important extension of the GLRT for multiple alternative hypotheses with differing numbers of parameters has been identified. Further work will be directed toward a comprehensive study of the theoretical properties of and applications of the EEF.

APPENDIX A. BACKGROUND ON EXPONENTIAL FAMILIES

The reader is referred to [1], [9], and [10] for further details. Consider the exponential pdf of (6). Some useful properties that we employ are now given.

1) The statistic $T(\mathbf{x})$ is a minimal and complete sufficient statistic for η. Hence, all the information in $\mathbf{x}$ can be reduced to $T(\mathbf{x})$ without degrading the performance of any detector/classifier based on it. We may therefore restrict our attention to $T(\mathbf{x})$ in distinguishing between $p(\mathbf{x};\eta_1)$ and $p(\mathbf{x};\eta_0)$ for any η_1 and η_0.

2) The moments of $T(\mathbf{x})$ are easily found from $K_0(\eta)$ or equivalently $M_0(\eta)$ since its moment generating function is

$$\begin{aligned}E_\eta(\exp(sT(\mathbf{x}))) &= \int \exp[(s+\eta)T(\mathbf{x}) - K_0(\eta)]p_0(\mathbf{x})d\mathbf{x}\\ &= E_0[\exp((s+\eta)T(\mathbf{x}))\exp(-K_0(\eta))]\\ &= M_0(s+\eta)\exp(-K_0(\eta)).\end{aligned}$$

As a result, it follows that

$$E_\eta(T) = M_0'(\eta)\exp(-K_0(\eta)) = \frac{M_0'(\eta)}{M_0(\eta)} = K_0'(\eta) \tag{38}$$

$$\begin{aligned}\mathrm{var}_\eta(T) &= \frac{M_0''(\eta)}{M_0(\eta)} - \left(\frac{M_0'(\eta)}{M_0(\eta)}\right)^2\\ &= K_0''(\eta).\end{aligned} \tag{39}$$

3) From (39) $K_0(\eta)$ is a convex function over $0 \le \eta \le 1$ since $K_0''(\eta) = \mathrm{var}_\eta(T) > 0$.

4) The MLE of η is easily found due to the convexity of $K_0(\eta)$. To do so we must find the η that maximizes $g(\eta) = \eta T(\mathbf{x}) - K_0(\eta)$. But $g(\eta)$ is concave and so we can differentiate and set equal to zero. The maximum will either be at this point or at one of the endpoints $\eta = 0$ or $\eta = 1$. Hence, we need to solve

$$T(\mathbf{x}) = K_0'(\eta)$$

for η to find the MLE $\hat{\eta}$. Also, the Fisher information for η is easily seen to be from (6) $I(\eta) = K_0''(\eta)$.

APPENDIX B. ASYMPTOTIC FORM OF REDUCED EEF

From (14) we have that

$$p(\mathbf{x};\eta) = \exp\left[\eta \ln\left(\frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p(\mathbf{x};\boldsymbol{\theta}_0)}\right) - K_0(\eta) + \ln p(\mathbf{x};\boldsymbol{\theta}_0)\right]$$

where $p_0(\mathbf{x})$ is now written as $p(\mathbf{x};\boldsymbol{\theta}_0)$. But asymptotically the MLE $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is a sufficient statistic with the asymptotic pdf $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}_1, \mathbf{I}^{-1}(\boldsymbol{\theta}_1))$, where $\mathbf{I}(\boldsymbol{\theta}_1)$ is the Fisher information matrix evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. Hence, we have that upon using the definition of conditional pdf

$$\ln\left(\frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p(\mathbf{x};\boldsymbol{\theta}_0)}\right) = \ln\left(\frac{p_1(\mathbf{x};\boldsymbol{\theta})}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})p(\mathbf{x};\boldsymbol{\theta}_0)}\right)$$

and since the conditional pdf cannot depend on $\boldsymbol{\theta}$, we let $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ to yield (with $p_1(\mathbf{x};\boldsymbol{\theta}_0) = p(\mathbf{x};\boldsymbol{\theta}_0)$)

$$\ln\left(\frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p(\mathbf{x};\boldsymbol{\theta}_0)}\right) = -\ln(p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)).$$

Now since $\mathbf{T} = \hat{\boldsymbol{\theta}}$ we have for $\boldsymbol{\theta}$ a $p \times 1$ vector

$$\begin{aligned}&p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)\\ &\quad = \underbrace{\frac{1}{(2\pi)^{(p/2)}\det^{1/2}(\mathbf{I}^{-1}(\boldsymbol{\theta}_0))}}_{c}\exp\left[-\frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right].\end{aligned}$$

Also, since $K_0(\eta)$ is the cumulant generating function of $-\ln(p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0))$, it will absorb the constant term c. Thus,

$$\ln\left(\frac{p_1(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}))}{p(\mathbf{x};\boldsymbol{\theta}_0)}\right) = \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

But when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is a chi-squared random variable with p degrees of freedom (denoted by χ_p^2) and thus

$$\begin{aligned}K_0(\eta) &= \ln E_0[\exp((\eta/2)\chi_p^2)]\\ &= \ln\frac{1}{(1-\eta)^{p/2}}.\end{aligned}$$

Finally, then the asymptotic form of the reduced EEF is

$$\begin{aligned}p(\mathbf{x};\eta) = \exp[&(\eta/2)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\\ &+ (p/2)\ln(1-\eta) + \ln p(\mathbf{x};\boldsymbol{\theta}_0)].\end{aligned}$$

This can be rewritten in terms of the GLRT by noting that asymptotically

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}})}{p(\mathbf{x};\boldsymbol{\theta}_0)}$$

which follows from the equivalence of the Wald test to the GLRT [11]. We let

$$l_G(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}})}{p(\mathbf{x};\boldsymbol{\theta}_0)}$$

be the GLRT statistic so that finally we have

$$p(\mathbf{x};\eta) = \exp[(\eta/2)l_G(\mathbf{x}) + (p/2)\ln(1-\eta) + \ln p(\mathbf{x};\boldsymbol{\theta}_0)] \tag{40}$$

which is the final form.

APPENDIX C. RATIONALE FOR MODEL ORDER ESTIMATOR

We define the divergence as [5]

$$D(p\,\|\,p_0) = \int p(\mathbf{x})\ln\frac{p(\mathbf{x})}{p_0(\mathbf{x})}d\mathbf{x} = E_p(\ln p/p_0).$$

We first prove that if a third pdf p_η has the property that

$$E_{p_\eta}(\ln p_\eta(\mathbf{x})/p_0(\mathbf{x})) = E_p(\ln p_\eta(\mathbf{x})/p_0(\mathbf{x})) \tag{41}$$

then a Pythagorean-like theorem holds [3]

$$D(p\,\|\,p_0) = D(p\,\|\,p_\eta) + D(p_\eta\,\|\,p_0).$$

Then, we argue that asymptotically this relationship will hold. A more rigorous proof can be found in [20]. Since

$$D(p\,\|\,p_\eta) + D(p_\eta\,\|\,p_0) = E_p(\ln p/p_\eta) + E_{p_\eta}(\ln p_\eta/p_0)$$

if (41) holds, then

$$\begin{aligned} D(p\,\|\,p_\eta) + D(p_\eta\,\|\,p_0) &= E_p(\ln p/p_\eta) + E_{p_\eta}(\ln p_\eta/p_0) \\ &= E_p(\ln p/p_\eta) + E_p(\ln p_\eta/p_0) \\ &= E_p(\ln p/p_0) \\ &= D(p\,\|\,p_0). \end{aligned}$$

Now consider p as the true model pdf, i.e., denote it by p_T. For the Pythagorean theorem to hold we require

$$E_{p_\eta}(\ln p_\eta(\mathbf{x})/p_0(\mathbf{x})) = E_{p_T}(\ln p_\eta(\mathbf{x})/p_0(\mathbf{x})). \tag{42}$$

But by the large of large numbers if $\mathbf{x}$ is composed of N IID random variables, then

$$\frac{1}{N}\ln p_\eta(\mathbf{x})/p_0(\mathbf{x})$$

will converge to its true expected value or

$$\begin{aligned} \frac{1}{N}\ln p_\eta(\mathbf{x})/p_0(\mathbf{x}) &= \frac{1}{N}\sum_{n=0}^{N-1}\ln\frac{p_\eta(x[n])}{p_0(x[n])} \\ &\to E\left[\ln\frac{p_\eta(x[n])}{p_0(x[n])}\right] = \frac{1}{N}E\left[\frac{\ln p_\eta(\mathbf{x})}{p_0(\mathbf{x})}\right]. \end{aligned}$$

Hence, we can replace (42) by the requirement

$$E_{p_\eta}(\ln p_\eta(\mathbf{x})/p_0(\mathbf{x})) = \ln p_\eta(\mathbf{x})/p_0(\mathbf{x}) \tag{43}$$

and note that if

$$p_\eta(\mathbf{x}) = p(\mathbf{x};\eta) = \exp[\eta T(\mathbf{x}) - K_0(\eta) + \ln p_0(\mathbf{x})]$$

then

$$\ln p_\eta(\mathbf{x})/p_0(\mathbf{x}) = \eta T(\mathbf{x}) - K_0(\eta)$$

and (43) becomes

$$E_{p_\eta}(\eta T(\mathbf{x}) - K_0(\eta)) = \eta T(\mathbf{x}) - K_0(\eta)$$

so that the requirement is just

$$E_{p_\eta}(T(\mathbf{x})) = T(\mathbf{x}). \tag{44}$$

But $E_{p_\eta}(T(\mathbf{x})) = K_0'(\eta)$ so (44) becomes $K_0'(\eta) = T(\mathbf{x})$, which is satisfied if η is chosen as the MLE of η. Hence, we have asymptotically the Pythagorean theorem

$$D(p_T\,\|\,p_0) = D(p_T\,\|\,p_{\hat{\eta}}) + D(p_{\hat{\eta}}\,\|\,p_0) \tag{45}$$

where $p_{\hat{\eta}} = \max_\eta p(\mathbf{x};\eta)$.

Now referring to (45) we see that since $D(p_T\,\|\,p_0)$ is fixed, if we wish to minimize the divergence between the true model and our estimated model, which is given by $D(p_T\,\|\,p_{\hat{\eta}})$, we should maximize the divergence between the estimated model and p_0 or $D(p_{\hat{\eta}}\,\|\,p_0)$. Since

$$\begin{aligned} D(p_{\hat{\eta}}\,\|\,p_0) &= E_{p_{\hat{\eta}}}\left(\ln\frac{p(\mathbf{x};\hat{\eta})}{p_0(\mathbf{x})}\right) \\ &= \hat{\eta}T(\mathbf{x}) - K_0(\hat{\eta}) \end{aligned}$$

we have

$$\mathrm{EEF}(i) = \max_\eta(\eta T_i(\mathbf{x}) - K_0(\eta))$$

which when applied to the reduced EEF is (27).

APPENDIX D. DETERMINATION OF ASYMPTOTIC MODEL ORDER ESTIMATOR

From (40) we have that

$$\mathrm{EEF}(i) = \max_\eta((\eta/2)l_{G_i}(\mathbf{x}) + (i/2)\ln(1-\eta)). \tag{46}$$

The MLE of η is easily shown to be

$$\hat{\eta} = \max\left(0, 1 - \frac{i}{l_{G_i}(\mathbf{x})}\right)$$

and substituting this value into (46) and multiplying by two we have that for $\hat{\eta} > 0$

$$\mathrm{EEF}(i) = l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right].$$

Finally we have that

$$\begin{aligned} \mathrm{EEF}(i) &= \begin{cases} l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right] & \frac{l_{G_i}(\mathbf{x})}{i} > 1 \\ 0 & \frac{l_{G_i}(\mathbf{x})}{i} \le 1. \end{cases} \\ &= \left(l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right]\right)u\left(\frac{l_{G_i}(\mathbf{x})}{i} - 1\right) \end{aligned}$$

where $u(x)$ is the unit step function. Thus, we have (32).

APPENDIX E. DERIVATION OF ASYMPTOTIC AR MODEL ORDER ESTIMATOR

To explicitly evaluate (33) we need the MLE of the AR parameters for the ith order model. Based on (34) the MLE is given as [19]

$$\hat{\mathbf{a}}_i = -(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T\mathbf{x}_i$$

where

$$\hat{\mathbf{a}}_i = [\hat{a}[1]\,\hat{a}[2]\ldots\hat{a}[i]]^T$$

$$\mathbf{x}_i = [x[i]\,x[i+1]\ldots x[N-1]]^T$$

$$\mathbf{H}_i = \begin{bmatrix} x[i-1] & x[i-2] & \ldots & x[0] \\ x[i] & x[i-1] & \ldots & x[1] \\ \vdots & \vdots & \vdots & \vdots \\ x[N-2] & x[N-3] & \ldots & x[N-1-i] \end{bmatrix}.$$

Also, we have that

$$\hat{\sigma}_{u_i}^2 = \frac{1}{N-i}\sum_{n=i}^{N-1}\left(x[n] + \sum_{k=1}^{i}\hat{a}_i[k]x[n-k]\right)^2$$

$$= \frac{1}{N-i}\mathbf{x}_i^T(\mathbf{I} - \mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T\mathbf{x}_i.$$

Thus, substituting into (34) produces

$$p(\mathbf{x};\hat{\mathbf{a}}_i,\hat{\sigma}_{u_i}^2) = \frac{1}{(2\pi\hat{\sigma}_{u_i}^2)^{(N-i)/2}}\exp[-(N-i)/2].$$

The constrained MLE for $\mathbf{a} = 0$ is obtained by maximizing (35) over σ_u^2. This is easily shown to produce

$$\hat{\sigma}_{u_0}^2 = \frac{1}{N-i}\sum_{n=i}^{N-1}x^2[n]$$

$$= \frac{1}{N-i}\mathbf{x}_i^T\mathbf{x}_i.$$

Substituting into (33) produces

$$p(\mathbf{x};\mathbf{a}=0,\hat{\sigma}_{u_0}^2) = \frac{1}{(2\pi\hat{\sigma}_{u_0}^2)^{(N-i)/2}}\exp[-(N-i)/2].$$

Therefore,

$$\frac{p(\mathbf{x};\hat{\mathbf{a}}_i,\hat{\sigma}_{u_i}^2)}{p(\mathbf{x};\mathbf{a}=0,\hat{\sigma}_{u_0}^2)} = \left(\frac{\hat{\sigma}_{u_0}^2}{\hat{\sigma}_{u_i}^2}\right)^{(N-i)/2}$$

and finally

$$l_{G_i}(\mathbf{x}) = (N-i)\ln\left(\frac{\hat{\sigma}_{u_0}^2}{\hat{\sigma}_{u_i}^2}\right).$$

REFERENCES

[1] Brown, L. D.
Fundamentals of Statistical Exponential Familes.
Institute of Mathematical Statistics, Monograph Series, 1986.

[2] Atkinson, A. C.
A method for discriminating between models.
Journal of Royal Statistical Society, **32** (1970), 323–353.

[3] Amari, S., and Nagaoka, H.
Methods of Information Geometry.
New York: Oxford, 1993.

[4] Murray, M. K., and Rice, J. W.
Differential Geometry and Statistics.
New York: Chapman and Hall, 1993.

[5] Kullback, S.
Information Theory and Statistics.
New York, Dover, 1959.

[6] Cover, T. M., and Thomas, J. A.
Elements of Information Theory.
New York: Wiley, 1991.

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

[8] Cox, D. R., and Hinkley, D. V.
Problems and Solutions in Theoretical Statistics.
New York: Chapman and Hall, 1978, 13.

[9] Barndorff-Nielsen
Information and Exponential Families.
New York: Wiley, 1978.

[10] Lehmann, E. L.
Testing Statistical Hypotheses, (2nd ed.).
New York: Springer-Verlag, 1986.

[11] Kay, S. M.
Fundamentals of Statistical Signal Processing: Detection Theory.
Englewood Cliffs, NJ: Prentice-Hall, 1998.

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

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

[14] Djuric, P.
Asymptotic MAP criteria for model selection.
IEEE Transactions on Signal Processing, (Oct. 1998), 2726–2735.

[15] Kay, S.
Conditional model order estimation.
IEEE Transactions on Signal Processing, (Sept. 2001), 1910–1917.

[16] Rissanen, J.
Modeling by shortest data description.
Automatica, (1978), 465–478.

[17] Kashyap, R. L.
Optimal choise of AR and MA parts in autoregressive moving average model.
IEEE Transactions on Pattern Analysis and Machine Intelligence, (1982), 99–104.

[18] Rao, C. R.
Linear Statistical Inference and its Applications.
New York: Wiley, 1973.

[19] Kay, S.
Modern Spectral Estimation: Theory and Application.
Englewood Cliffs, NJ: Prentice-Hall, 1988.

[20] White, H.
Estimation, Inference, and Specification Analysis.
New York: Cambridge, 1996.

Steven Kay (M’75—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, 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. His current interests are spectrum analysis, detection and estimation theory, and statistical signal processing.

Dr. Kay has written numerous papers and is a contributor to several edited books. He is the author of the textbooks *Modern Spectral Estimation* (Prentice-Hall, 1988), *Fundamentals of Statistical Signal Processing, Vol. I: Estimation Theory* (Prentice-Hall, 1993), and *Fundamentals of Statistical Signal Processing, Vol. II: Detection Theory* (Prentice-Hall, 1998). He is a member of Tau Beta Pi and Sigma Xi. He has been a distinguished lecturer for the IEEE Signal Processing Society. He has served on the IEEE Acoustics, Speech, and Signal Processing Committee on Spectral Estimation and Modeling, on IEEE Oceans committees, and is currently an associate editor for the *IEEE Signal Processing Letters*. Dr. Kay has recently been included on a list of the 250 most cited researchers in engineering in the world.