The Estimation of the Model Order in Exponential Families

NERI MERHAV, MEMBER, IEEE

Abstract—**The estimation of the model order in exponential families is studied. Estimators are sought that achieve high exponential rate of decrease in the underestimation probability while keeping the overestimation probability exponent at a certain prescribed level. It is assumed that a given integer is known to upper-bound the true order.**

I. Introduction

The problem of estimating the order of a statistical model has been studied in the literature of time-series analysis, information theory and automatic control. Most of the estimators proposed (Akaike [1]–[3], Kayshap [4], Shibata [5]–[8], Rissanen [9]–[12], Parzen [13], Hannan [14], Hanna and Quinn [15], Schwarz [16], Tong [17], Wax and Kailath [18], Broersen [19], and others) are heuristic in character, and no optimality results concerning the order estimation error (stronger than consistency) have been established. An exception is Schwarz [16] who, in fact, proved the optimality of the minimum description length (MDL) principle (Rissanen [9]–[12]) in a Bayesian sense. However, it should be pointed out that Rissanen's results are motivated by coding applications, as they are not limited to the case where a "true" order does exist [12]. (A more detailed discussion about the links between the present method and Rissanen's approach appears in [20].) Among other works where it is not assumed that a true finite order exists are Shibata [6], [7] in which an infinite order autoregressive (AR) process is assumed. In [6] Shibata proposes an order estimator that is optimal in the sense of minimizing the mean squared error of the estimated predictor. In [7] Shibata demonstrates that the same order estimator is also optimal in the sense of minimizing the integrated relative spectral squared error. In [8] Shibata studies the relationship between consistency of model selection and that of parameter estimation.

This correspondence is an extension of an earlier paper [20] in which the estimation of the order of a discrete finite Markov chain was studied. In [20] we derived order estimators $\hat{k}$ having the smallest underestimation probability $\Pr(\hat{k} < k)$ among all universal estimators for which the overestimation probability $\Pr(\hat{k} > k)$ decays faster than $2^{-\lambda n}$ for a given value of $\lambda > 0$, where n is the sample size. Here, we use the same performance criterion in a more general situation, where the observations $x_1, \cdots, x_n$ (taking continuous real values) emerge from a source of the exponential family. In contrast to Rissanen [9]–[12] and Schwarz [16], we are able to attain an exponentially fast vanishing error probability, as we are not using the Bayesian formulation.

The outline of the correspondence is as follows. In Section II we formulate the problem. The analysis and main results are presented in Section III. Finally, in Sections IV and V, some examples are given for possible applications of the proposed method in specific order estimation and hypothesis testing problems.

Manuscript received January 14, 1988; revised August 2, 1988.

The author was with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. He is now with AT&T Bell Laboratories, Room 2C-543, 600 Mountain Avenue, Murray Hill, NJ 07974.

IEEE Log Number 8930782.

II. Problem Formulation

Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a vector of independent identically distributed (i.i.d.) real-valued random variables governed by a probability distribution from the k-parameter exponential family (Koopman–Darmois)

$$dP_{\theta^k}(\boldsymbol{x}) = \exp\left[\theta^k \circ T^k(\boldsymbol{x}) - \psi(\theta^k)\right]\mu(dx) \tag{1}$$

where $\circ$ denotes the scalar product of two vectors, θ^k takes values in a bounded open subset Θ_k of $\mathbb{R}^k$, for which $\int_{\mathbb{R}} \exp[\theta^k \circ T^k(x)]\mu(dx) < \infty$, with $T^k(x)$ a k-dimensional real-valued statistic and $\mu(\cdot)$ is a σ-finite measure on Borel subsets of $\mathbb{R}$. The function $\psi(\theta^k)$ is chosen to normalize (1), namely,

$$\psi(\theta^k) = \log \int_{-\infty}^{\infty} \exp\left[\theta^k \circ T^k(x)\right]\mu(dx). \tag{2}$$

$\psi(\theta^k)$ is called the log moment generating function, as it yields the moments of P_{θ^k} by differentiation with respect to $\theta^k \in \Theta_k$. Since the x_i are i.i.d., clearly,

$$dP_{\theta^k}(\boldsymbol{x}) = \exp\left\{ n\left[\theta^k \circ \frac{1}{n}\sum_{i=1}^{n} T^k(x_i) - \psi(\theta^k)\right]\right\} \prod_{i=1}^{n} \mu(dx_i). \tag{3}$$

The aim of this correspondence is to derive an order estimator $\hat{k}$, that minimizes the underestimation probability $P_{\theta^k}(\hat{k} < k)$, uniformly for *every* $\theta^k \in \Theta_k$, subject to the constraint

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log P_{\theta^k}(\hat{k} > k)\right] > \lambda, \qquad \forall 1 \le k < k_0, \forall \theta^k \in \Theta_k. \tag{4}$$

It is assumed that a given positive integer k_0 is larger than the true order k. (A similar performance criterion was introduced in [20].)

III. Main Results

We assume the following regularity conditions.

1) For $1 \le j \le k_0$, the parameter space Θ_j is a bounded open subset of

$$\eta_j \triangleq \left\{\theta^j : \int_{-\infty}^{\infty} e^{\theta^j \circ T^j(x)}\mu(dx) < \infty\right\}.$$

2) The set of equations:

$$\nabla\psi(\theta^j) = \frac{1}{n}\sum_{i=1}^{n} T^j(x_i)$$

has a unique solution $\hat{\theta}^j_{\mathrm{ML}} \in \eta_j$ for any $\boldsymbol{x} \in \mathbb{R}^n$, $1 \le j \le k_0$.

3) For any $\theta^j \in \Theta_j$, the Fisher information matrix

$$\nabla^2\psi(\theta^j) \triangleq \left\{\partial^2\psi(\theta^j)\Big/\partial\theta^j_p\,\partial\theta^j_q\right\}_{p,q=1}^{j} = \mathrm{cov}_\theta\left\{T^j(x)\right\}$$

is positive-definite and bounded.

Denote by $\hat{\theta}^j_{\mathrm{ML}}(\boldsymbol{x})$, the maximum likelihood estimator of θ^j. Clearly, by condition 2) $\hat{\theta}^j_{\mathrm{ML}}$ satisfies

$$\nabla\psi\left(\hat{\theta}^j_{\mathrm{ML}}\right) \equiv \frac{1}{n}\sum_{i=1}^{n} T^j(x_i). \tag{5}$$

Let us denote by $A^n_\epsilon(\theta^j)$, $\theta^j \in \Theta_j$ the set of vectors $\boldsymbol{x} \in \mathbb{R}^n$ for which

$$\left\|\frac{1}{n}\sum_{i=1}^{n} T^j(x_i) - \nabla\psi(\theta^j)\right\|_j < \epsilon,$$

0018-9448/89/0900-1109$01.00 ©1989 IEEE

where $\|\cdot\|_j$ denotes the Euclidean norm in $\mathbf{R}^j$. That is, $A_\epsilon^n(\theta^j)$ denotes a "neighborhood" of vectors x whose maximum likelihood estimators $\hat{\theta}_{\mathrm{ML}}^j(x)$ are close to θ^j (in the sense just defined). This can be viewed as an extension to the definition of typical sequences [21] in the discrete alphabet case. (See also Rissanen [22, appendix A].)

We first prove an auxiliary lemma.

Lemma 1: For any $\delta > 0$, there exist $\epsilon > 0$ and n sufficiently large, such that the μ-measure of $A_\epsilon^n(\theta^j)$ is bounded as follows:

$$\int_{A_\epsilon^n(\theta^j)} \prod_{i=1}^n \mu(dx_i) \geq \exp\left\{-n\left[\theta^j \circ \nabla\psi(\theta^j) - \psi(\theta^j) + \delta\right]\right\}, \quad j = 1, 2, \cdots, k_0. \tag{6}$$

Proof: By condition 3) it follows that

$$\frac{1}{n}\sum_{i=1}^n T^j(x_i) \xrightarrow{P_{\theta^j}} \nabla\psi(\theta^j)$$

for any $\theta^j \in \Theta_j$. Thus, for any $\delta > 0$, $\epsilon > 0$, and n sufficiently large,

$$P_{\theta^j}\left\{\left\|\frac{1}{n}\sum_{i=1}^n T^j(x_i) - \nabla\psi(\theta^j)\right\|_j \leq \epsilon\right\} \geq 1 - \delta. \tag{7}$$

Thus

$$\begin{aligned}
1 - \delta &\leq \int_{x \in A_\epsilon^n(\theta^j)} \exp\left\{n\left[\theta^j \circ \frac{1}{n}\sum_{i=1}^n T^j(x_i) - \psi(\theta^j)\right]\right\} \cdot \prod_{i=1}^n \mu(dx_i) \\
&= \int_{x \in A_\epsilon^n(\theta^j)} \exp\left\{n\left[\theta^j \circ \left(\nabla\psi(\theta^j) + \frac{1}{n}\sum_{i=1}^n T^j(x_i) - \nabla\psi(\theta^j)\right) - \psi(\theta^j)\right]\right\} \prod_{i=1}^n \mu(dx_i)
\end{aligned}$$

(by the Schwarz-Cauchy inequality)

$$\begin{aligned}
&\leq \int_{x \in A_\epsilon^n(\theta^j)} \exp\left\{n\left[\theta^j \circ \nabla\psi(\theta^j) + \epsilon\|\theta^j\|_j - \psi(\theta^j)\right]\right\} \prod_{i=1}^n \mu(dx_i) \\
&= \exp\left\{n\left[\theta^j \circ \nabla\psi(\theta^j) + \epsilon\|\theta^j\|_j - \psi(\theta^j)\right]\right\} \cdot \int_{x \in A_\epsilon^n(\theta^j)} \prod_{i=1}^n \mu(dx_i) \\
&\leq \exp\left\{n\left[\theta^j \circ \nabla\psi(\theta^j) - \psi(\theta^j) + \delta\right]\right\} \cdot \int_{x \in A_\epsilon^n(\theta^j)} \prod_{i=1}^n \mu(dx_i).
\end{aligned} \tag{8}$$

This completes the proof.

Assuming that a given positive integer k_0 is known to upper-bound the true order k, define the following order estimator:

$$k^* = \min\left\{j: \frac{1}{n}\log\frac{dP_{\hat{\theta}_{\mathrm{ML}}^{k_0}}(x)}{dP_{\hat{\theta}_{\mathrm{ML}}^j}(x)} < \lambda\right\}, \tag{9}$$

where $dP_\theta(x)$ is defined as in (3). Thus (9) involves the Radon–Nikodym derivative computed at $\theta = \hat{\theta}_{\mathrm{ML}}^{k_0}$ and $\theta = \hat{\theta}_{\mathrm{ML}}^j$. Notice that the estimator k^* defined in (9) is a straightforward extension of the generalized likelihood ratio test (GLRT), which is widely used for composite hypotheses.

The algorithm starts from $j = 1$ and seeks the first integer j, for which $(1/n)\log dP_{\hat{\theta}_{\mathrm{ML}}^j}(x)$ is sufficiently close to $(1/n)\log dP_{\hat{\theta}_{\mathrm{ML}}^{k_0}}(x)$ (difference less than λ); that is, we find the integer j, for which any increase in the order will not significantly increase the likelihood.

Notice that, since k^* is a stopping rule, the associated computational complexity is usually smaller than for other existing order estimators (e.g., AIC, BIC, CAT, MDL, FPE, etc.), which in turn minimize a certain information criterion $\mathrm{IC}(j)$ over the integers $1 \leq j \leq k_0$. In fact, (9) is asymptotically equivalent to $\min\{j: \mathrm{IC}(j) - \mathrm{IC}(k_0) < \lambda\}$, where IC denotes any one of the above mentioned information criteria (AIC, BIC, etc.). The following theorem establishes the optimality of k^* in the sense defined in Section II.

Theorem 1: Assuming conditions 1)–3) are met, we have

a)

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log P_{\theta^k}(k^* > k)\right] \geq \lambda, \qquad \forall 1 \leq k < k_0, \forall \theta^k \in \Theta_k;$$

b) for any estimator $\hat{k}$ that satisfies (4), every $\theta^k \in \Theta_k$, $1 < k \leq k_0$ and n sufficiently large,

$$P_{\theta^k}(k^* < k) \leq P_{\theta^k}(\hat{k} < k).$$

Proof: Define

$$M_j \triangleq \left\{x: \frac{1}{n}\log\frac{dP_{\hat{\theta}_{\mathrm{ML}}^{k_0}}(x)}{dP_{\hat{\theta}_{\mathrm{ML}}^j}(x)} < \lambda\right\} \tag{10}$$

and the Kullback–Leibler information:

$$\begin{aligned}
K(\theta^j\|\varphi^j) &\triangleq E_{\theta^j}\left[\frac{dP_{\theta^j}(x)}{dP_{\varphi^j}(x)}\right] \\
&= (\theta^j - \varphi^j) \circ \nabla\psi(\theta^j) - \psi(\theta^j) + \psi(\varphi^j), \\
&\qquad \theta^j, \varphi^j \in \Theta_j.
\end{aligned} \tag{11}$$

This definition can be extended to the case where the dimensions of θ and φ are not necessarily equal by padding the lower dimension parameter vector with zeros.

As for part a), if follows by (3), (5), and (11) that $K(\hat{\theta}_{\mathrm{ML}}^{k_0}\|\theta^k) = (1/n)\log(dP_{\hat{\theta}_{\mathrm{ML}}^{k_0}}(x)/dP_{\theta^k}(x))$. Therefore,

$$\begin{aligned}
P_{\theta^k}(k^* > k) &= P_{\theta^k}\left\{\bigcap_{j \leq k} M_j^c\right\} \leq P_{\theta^k}\{M_k^c\} \\
&= P_{\theta^k}\left\{\frac{1}{n}\log\frac{dP_{\hat{\theta}_{\mathrm{ML}}^{k_0}}(x)}{dP_{\hat{\theta}_{\mathrm{ML}}^k}(x)} \geq \lambda\right\} \\
&\leq P_{\theta^k}\left\{\frac{1}{n}\log\frac{dP_{\hat{\theta}_{\mathrm{ML}}^{k_0}}(x)}{dP_{\theta^k}(x)} \geq \lambda\right\} \\
&= P_{\theta^k}\left\{K\left(\hat{\theta}_{\mathrm{ML}}^{k_0}\|\theta^k\right) \geq \lambda\right\} \leq e^{-(\lambda-\epsilon)n},
\end{aligned} \tag{12}$$

for any $\epsilon > 0$ and n large enough. The last inequality follows from a known result from the theory of large deviations for exponential families [23, theorem 6], namely

$$P_{\theta^k}\left\{K\left(\hat{\theta}_{\mathrm{ML}}^{k_0}\|\theta^k\right) \geq \lambda\right\} = n^{(k_0-2)/2}e^{-\lambda n}C_\lambda\left[1 + o_n(1)\right],$$

uniformly in λ over the range

$$\epsilon \leq \lambda \leq \sup\left\{\lambda: \left\{\theta: K(\theta\|\theta^k) \leq \lambda\right\} \subseteq \Theta_{k_0}\right\} - \epsilon$$

where C_λ is a constant not depending on n. This completes the proof of part a).

To prove b), let $\hat{k}$ be an arbitrary order estimator satisfying (4). Let $\Omega_j = \{x \in \mathbb{R}^n: \hat{k} = j\}$, $j = 1, \cdots, k_0$. Clearly, $\{\Omega_j\}_{j=1}^{k_0}$ is a partition of $\mathbb{R}^n$. We also assume that for any $\theta^{k_0} \in \eta_{k_0}$, we have $A_\epsilon^n(\theta^{k_0}) \subseteq \Omega_j$ for some $1 \le j \le k_0$, and $\epsilon > 0$ sufficiently small. In other words, $(1/n)\sum_{i=1}^n T^{k_0}(x_i)$ is a sufficient statistic for $\{\Omega_j\}_{j=1}^{k_0}$ (this assumption does not affect the asymptotic exponent associated with $P_{\theta^k}(\hat{k} < k)$).

Let $x' \in \bigcup_{j>k} \Omega_j$ be an arbitrary n-tuple. Thus, by (4), for any $1 \le r \le k$, $\delta > 0$, and n sufficiently large,

$$
\begin{aligned}
e^{-(\lambda+\delta)n} &\ge \max_{\theta^r \in \Theta_r} P_{\theta^r}(\hat{k} > k) \\
&= \sum_{j \ge k} \int_{x \in \Omega_j} \exp\left\{ n\left[\hat{\theta}^r_{\mathrm{ML}}(x') \circ \frac{1}{n}\sum_{i=1}^n T^r(x_i) \right.\right. \\
&\qquad \left.\left. - \psi\left(\hat{\theta}^r_{\mathrm{ML}}(x')\right)\right]\right\} \prod_{i=1}^n \mu(dx_i) \\
&\ge \int_{x \in A_\epsilon^n\left(\hat{\theta}^{k_0}_{\mathrm{ML}}(x')\right)} \exp\left\{ n\left[\hat{\theta}^r_{\mathrm{ML}}(x') \circ \frac{1}{n}\sum_{i=1}^n T^r(x_i) \right.\right. \\
&\qquad \left.\left. - \psi\left(\hat{\theta}^r_{\mathrm{ML}}(x')\right)\right]\right\} \prod_{i=1}^n \mu(dx_i) \\
&\ge \left[\int_{A_\epsilon^n\left(\hat{\theta}^{k_0}_{\mathrm{ML}}(x')\right)} \prod_{i=1}^n \mu(dx_i) \right] \min_{x \in A_\epsilon^n\left(\hat{\theta}^{k_0}_{\mathrm{ML}}(x')\right)} \\
&\qquad \cdot \exp\left\{ n\left[\hat{\theta}^r_{\mathrm{ML}}(x') \circ \frac{1}{n}\sum_{i=1}^n T^r(x_i) - \psi\left(\hat{\theta}^r_{\mathrm{ML}}(x')\right)\right]\right\}
\end{aligned}
$$

(by Lemma 1)

$$
\begin{aligned}
&\ge \exp\left\{ -n\left[\hat{\theta}^{k_0}_{\mathrm{ML}}(x') \circ \frac{1}{n}\sum_{i=1}^n T^{k_0}(x_i') - \psi\left(\hat{\theta}^{k_0}_{\mathrm{ML}}(x')\right) + \frac{\delta}{2}\right]\right\} \\
&\qquad \cdot \exp\left\{ n\left[\hat{\theta}^r_{\mathrm{ML}}(x') \circ \frac{1}{n}\sum_{i=1}^n T^r(x_i') - \psi\left(\hat{\theta}^r_{\mathrm{ML}}(x')\right) - \frac{\delta}{2}\right]\right\} \\
&= \exp\left\{ -n\left[\frac{1}{n}\log \frac{dP_{\hat{\theta}^{k_0}_{\mathrm{ML}}}(x')}{dP_{\hat{\theta}^r_{\mathrm{ML}}}(x')} + \delta\right]\right\}. \qquad (13)
\end{aligned}
$$

It now follows from (13) that

$$
\bigcup_{j>k} \Omega_j \subseteq \bigcap_{r \le k} M_r^c, \qquad 1 \le k \le k_0 - 1 \qquad (14)
$$

(where the superscript c denotes the complement set). Equivalently,

$$
\bigcup_{j \le k} M_j \subseteq \bigcap_{j>k} \Omega_j^c = \bigcup_{j \le k} \Omega_j. \qquad (15)
$$

Hence

$$
P_{\theta^k}(k^* < k) = P_{\theta^k}\left(\bigcup_{j \le k-1} M_j \right) \le P_{\theta^k}\left(\bigcup_{j \le k-1} \Omega_j \right) = P_{\theta^k}(\hat{k} < k). \qquad (16)
$$

This completes the proof of Theorem 1. □

As mentioned in [20, remark 1], the value of λ should be chosen sufficiently small to guarantee an exponential decay of *both* overestimation and underestimation probabilities. This is different from other existing estimators (AIC, BIC, MDL) where the overestimation probability can be shown [20] to decay slower than $2^{-\epsilon n}$ for any $\epsilon > 0$.

IV. Examples

In this section the algorithm k^* defined in (9) is applied to several well-known models.

A. The Linear Regression Model

Let

$$
y_i = \sum_{l=1}^k a_l x_i^l + w_i, \qquad i = 1, \cdots, n \qquad (17)
$$

where $\{w_i\}$ are i.i.d. Gaussian zero-mean random variables with unknown variance σ^2 and $x_j = (x_1^j, x_2^j, \cdots, x_n^j)$, $j = 1, 2, \cdots, k_0$ ($k \le k_0 < n$) are given linearly independent vectors. Without loss of generality, let us assume that the x_j are orthonormal, namely $(1/n)\sum_{i=1}^n x_i^j x_i^l = \delta_{lj}$, where δ_{lj} is the indicator function for $j = l$. We are interested in estimating k. In what follows, we first demonstrate that this is a special case of the exponential family (3). We let $a^k = (a_1, \cdots, a_k)$, and $y = (y_1, \cdots, y_n)$:

$$
\begin{aligned}
&dP_{a^k}(y) \Big/ \prod_{i=1}^n dy_i \\
&\quad = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \sum_{l=1}^k a_l x_i^l \right)^2 \right\} \\
&\quad = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \sum_{l=1}^k \frac{a_l}{\sigma^2}\sum_{i=1}^n x_i^l y_i \right. \\
&\qquad \left. - n\left[\sum_{l=1}^k \frac{a_l^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2 \right]\right\}. \qquad (18)
\end{aligned}
$$

Clearly, (18) agrees with (3), where

$$
\theta^{k+1} = (\theta_1, \cdots, \theta_{k+1}) = \left(-\frac{1}{2\sigma^2}, \frac{a_1}{\sigma^2}, \cdots, \frac{a_k}{\sigma^2} \right), \qquad (19)
$$

$$
T^{k+1}(y_i) = \left(y_i^2, x_i^1 y_i, \cdots, x_i^k y_i \right), \qquad (20)
$$

$$
\begin{aligned}
\psi(\theta^{k+1}) &= \sum_{l=1}^k \frac{a_l^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2 \\
&= -\frac{1}{4\theta_1}\sum_{i=2}^{k+1} \theta_i^2 + \frac{1}{2}\log\left(-\frac{1}{2\theta_1} \right), \qquad (21)
\end{aligned}
$$

and $\mu(\cdot)$ is the Lebesgue measure.

Applying (9) to this case we straightforwardly obtain

$$
k^* = \min\left\{ j: \frac{1}{2}\log\hat{\sigma}_j^2 - \frac{1}{2}\log\hat{\sigma}_{k_0}^2 < \lambda \right\} \qquad (22)
$$

where $\hat{\sigma}_j^2$ is the minimum residual energy of order j given by

$$
\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{l=1}^j \hat{a}_l x_i^l \right)^2, \qquad (23)
$$

with $(\hat{a}_1, \cdots, \hat{a}_j)$ the ML estimators of $(a_1, \cdots, a_j)$ assuming a jth-order model. Thus, if $\hat{\sigma}_j^2$ is sufficiently close to $\hat{\sigma}_{k_0}^2$ for some j, the algorithm stops. Notice also that (22) involves a difference between empirical entropies of orders j and k_0. In fact, this is the case in all of the following examples.

B. The Autoregressive Model

Although the observations produced by an AR model are not independent, the results of Section III can be shown to continue to hold. In fact, this case is similar to the previous example,

where each x_i^j is replaced by $-y_{i-j}$; we have

$$dP_{a^k}(y)\Big/\prod_{i=1}^{n} dy_i = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{l=1}^{n}\left(y_l+\sum_{j=1}^{k} a_j y_{l-j}\right)^2\right]. \quad (24)$$

We now put (24) in the standard exponential form (3) with

$$\theta^{k+1} = -\left(\frac{1}{2\sigma^2}, \frac{1}{2\sigma^2}R_a(1), \cdots, \frac{1}{2\sigma^2}R_a(k)\right) \quad (25)$$

where $R_a(i)$, $1 \le i \le k$, is the autocorrelation of the whitening filter $A(z) = 1+\sum_{i=1}^{k} a_i z^{-1}$, namely,

$$R_a(i) \triangleq \sum_{j=0}^{k-i} a_{i+j}a_j, \qquad a_0 = 1 \quad (26)$$

and

$$T^{k+1}(y_{i-1}, \cdots, y_{i-k}) = \left(y_i^2, y_i y_{i-1}, \cdots, y_i y_{i-k}\right) \quad (27)$$

$$\psi(\theta) = \frac{1}{2}\log\sigma^2 = \frac{1}{2}\log\left(-\frac{1}{2\theta_1}\right). \quad (28)$$

Notice that T^k depends now on y_i, $y_{i+1}, \cdots, y_{i-k}$. The main difference with respect to Example A, is that here $y_{i-j} = -x_i^l$ are not deterministic. However, we use the fact that $\sum_{i=1}^{n} y_{i-k}y_{i-k-j}$ is approximately independent of k, since ignoring the k edge observations does not affect the asymptotic result of Theorem 1. Again, $\mu(\cdot)$ is Lebesgue measure. Applying Theorem 1, we obtain the same estimator as (22), where now $\hat{\sigma}_j^2$ is the empirical prediction residual of order j, that is,

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i+\sum_{l=1}^{j}\hat{a}_l y_{i-j}\right)^2, \quad (29)$$

where $\{\hat{a}_l\}_{l=1}^{j}$ are the jth-order ML estimators of $\{a_l\}_{l=1}^{j}$ obtained by the Yule–Walker equations.

C. The Discrete Markov Source

Consider a discrete-alphabet Markov source of order k, that is,

$$P(x_i|x_{-\infty}^{i-1}) = P(x_i|x_{i-k}^{i-1}). \quad (30)$$

Suppose it is desired to estimate the order k. The vector x_{i-k}^{i-1} will be referred to as the state s_i. This problem was studied in detail in [20].

To demonstrate that the Markovian source is a member of the exponential family, we concentrate on the first-order case for simplicity.

Let $X = \{1,2,\cdots,m\}$ be the output alphabet of the source; then

$$P(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|x_{i-1}) = \prod_{u=1}^{m}\prod_{v=1}^{m} p(u|v)^{n(u,v)} = \exp\left[\sum_{u=1}^{m}\sum_{v=1}^{m} n(u,v)\log p(u|v)\right] \quad (31)$$

where $p(u|v)$ is the transition probability from $v \in X$ to $u \in X$, and $n(u,v)$ is the number of such transitions in the sequence $\boldsymbol{x}$. The sufficient statistics $\{n(u,v)\}$ are redundant since they satisfy:

$$\sum_{u=1}^{m} n(u,v) = \sum_{u=1}^{m} n(v,u) \pm 1. \quad (32)$$

The term ± 1 can be ignored as it does not affect the asymptotic exponent. Plugging (32) into (31), we get after some algebra,

$$P(\boldsymbol{x}) \cong \exp\left[\sum_{u=1}^{m-1}\sum_{v=1}^{m} n(u,v)\log\frac{p(u|v)\,p(m|u)}{p(m|v)\,p(m|m)} + n\log p(m|m)\right]. \quad (33)$$

Here $P(\boldsymbol{x})$ is expressed by $m(m-1)$ parameters. Therefore, we have the exponential form (3) with

$$\theta_{uv} = \log\left[\frac{p(u|v)\,p(m|u)}{p(m|v)\,p(m|m)}\right] \quad (34)$$

$$T^{uv}(x_i, x_{i-1}) = \delta(x_i = u,\ x_{i-1} = v) \quad (35)$$

for $u = 1,\cdots,m-1$ and $v = 1,\cdots,m$, where $\delta(x_i = u,\ x_{i-1} = v)$ is the indicator function for $x_i = u$ jointly with $x_{i-1} = v$.

The function $\psi(\theta)$ is given by

$$\psi(\theta) = -\log p(m|m), \quad (36)$$

and $\mu(\cdot)$ is the counting measure. The estimator k^* for this case can be shown [20] to be

$$k^* = \min\{j:\ H_j(\boldsymbol{x}) - H_{k_0}(\boldsymbol{x}) < \lambda\}, \quad (37)$$

where $H_j(\boldsymbol{x})$ is the jth order empirical conditional entropy of $\boldsymbol{x}$. Notice that again, as in the previous examples, the estimator k^* involves differences between empirical entropies.

V. Applications to Composite Binary Hypothesis Testing

Several well-known binary hypothesis testing problems can be formalized in the present framework. In these cases, we have only two hypotheses concerning the order of a statistical model. Our performance criterion will now coincide with the regular Neyman–Pearson criterion.

A. Testing for Independence

Suppose we are given n i.i.d. observed pairs (x_1, y_1), $(x_2, y_2), \cdots, (x_n, y_n)$ of random variables and it is desired to test the null hypothesis H_0: x and y are statistically independent, against the alternative H_1: x and y are dependent. Consider two examples:

1) The Gaussian Model: Let $x \sim N(0,\sigma^2)$, $y \sim N(0,\sigma^2)$ and

$$H_0:\ dP_\theta(x,y)/dx\,dy = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{x^2+y^2}{2\sigma^2}\right]$$

$$H_1:\ dP_\theta(x,y)/dx\,dy = \frac{1}{2\pi\sigma^2(1-\rho^2)^{1/2}}\exp\left[-\frac{x^2-2\rho xy+y^2}{2\sigma^2(1-\rho^2)}\right].$$

Notice that under hypothesis H_0 we have only one free parameter $\theta^1 = -1/2\sigma^2$, $k=1$, while under H_1 we have two free parameters ($k=2$): $\theta^2 = (-1/2\sigma^2(1-\rho^2)^{1/2},\ \rho/\sigma^2(1-\rho^2)^{1/2})$ The "estimator" k^* in this case turns out to be the following test. Reject H_0 iff

$$\frac{1}{2}\log\frac{1}{1-\hat{\rho}^2} > \lambda \quad (38)$$

where

$$\hat{\rho} = \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n}\left(x_i^2+y_i^2\right)}. \quad (39)$$

That is, the optimal test compares $|\hat{\rho}|$ to a certain threshold.

2) A Discrete Memoryless Model: Let x and y be random variables taking values in finite alphabets X and Y of sizes α and β, respectively. It is desired to test

$$H_0\colon P(x,y) = P(x)P(y) \text{ (independence)}$$

against

$$H_1\colon P(x,y) \neq P(x)P(y).$$

Clearly, under H_0 we have $k=(\alpha-1)(\beta-1)$ free parameters, whereas under H_1 we have $k=\alpha\beta-1$ parameters. The resulting test rejects H_0 iff

$$\hat{I}(x;y) > \lambda \tag{40}$$

where $\hat{I}(x;y)$ is the empirical mutual information between x and y. This test is optimal in the Neyman–Pearson sense. A similar result was presented by Gutman [24].

B. Testing for Equal Distributions

Suppose we are given two sequences of i.i.d. random variables $x_1,\cdots,x_n$ and $y_1,\cdots,y_m$ and wish to decide whether or not these two sequences were emitted from the same source. Consider three simple examples.

1) The Gaussian Model: Let $x_i \sim N(\mu_1,\sigma^2)$ and $y_j \sim N(\mu_2,\sigma^2)$; suppose we are testing $H_0\colon \mu_1=\mu_2=\mu$, against the alternative $H_1\colon \mu_1\neq\mu_2$. Under H_0, we have $k=2$ free parameters

$$\theta^2 = \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right),$$

while under H_1 there are $k=3$ parameters

$$\theta^3 = \left(-\frac{1}{2\sigma^2}, \frac{\mu_1}{\sigma^2}, \frac{\mu_2}{\sigma^2}\right).$$

The resulting test will reject H_0 iff

$$\frac{1}{2}\log\hat{\sigma}_0^2 - \frac{1}{2}\log\hat{\sigma}_1^2 > \lambda \tag{41}$$

with

$$\hat{\sigma}_0^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}(x_i-\bar{z})^2 + \sum_{j=1}^{m}(y_j-\bar{z})^2\right] \tag{42}$$

and

$$\hat{\sigma}_1^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{j=1}^{m}(y_j-\bar{y})^2\right] \tag{43}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m}y_i, \qquad \bar{z} = \frac{n\bar{x}+m\bar{y}}{n+m}.$$

That is, the variance is estimated under H_0 and under H_1, respectively. If the estimates are sufficiently close, then we decide in favor of H_0.

2) The One-Sided Exponential Model: Suppose that now x_i and y_j are distributed as follows:

$$dP_a(x)/dx = \begin{cases} ae^{-ax}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$dP_b(y)/dy = \begin{cases} be^{-by}, & y \geq 0 \\ 0, & y < 0 \end{cases}.$$

We want to decide whether $a=b$ or $a\neq b$ under the Neyman–Pearson criterion. Here we test $\theta^1=a$ against $\theta^2=(a,b)$. By (9) we reject H_0 iff

$$\log\bar{z} - \frac{n}{n+m}\log\bar{x} - \frac{m}{n+m}\log\bar{y} > \lambda, \tag{44}$$

where $\bar{x}$, $\bar{y}$, and $\bar{z}$ are defined as in example B-1. Again, the test statistic is a difference between empirical entropies. This result might be applicable to optical detection, where each observation denotes the difference in the time of arrivals of two successive photons. One of the sequences, say x, can be used as the training sequence, if the dark current parameter $a(a<b)$ is unknown, while the other plays the role of the test sequence to determine whether or not an optical signal is present.

3) The Discrete Memoryless Model: Let x and y be random variables as in A-2, where $X=Y$, $|X|=\alpha$; suppose we want to decide whether x and y were emitted from the same memoryless source. Under H_0, we have $k=\alpha-1$ free parameters, while under H_1, there are as many as $k=2(\alpha-1)$ parameters. The resulting test would reject H_0 iff

$$\left(1+\frac{m}{n}\right)\hat{H}(xy) - \hat{H}(x) - \frac{m}{n}\hat{H}(y) > \lambda \tag{45}$$

where $\hat{H}(x)$ and $\hat{H}(y)$ denote the empirical entropies of x and y, respectively, and $\hat{H}(xy)$ denotes the empirical entropy of the concatenation of x and y. A similar result was derived by Gutman [25].

Acknowledgment

Fruitful discussions with Professor Jacob Ziv are acknowledged. The author is also grateful to the anonymous referees for their helpful comments and suggestions.

References

[1] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, vol. 21, pp. 243–247, 1969.

[2] ____, "Statistical Predictor Identification," *Ann. Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

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

[4] R. L. Kayshap, "Inconsistency of the AIC rule for estimating the order of autoregressive models," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 996–998, 1980.

[5] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117–126, 1976.

[6] ____, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Ann. Statist.*, vol. 8, pp. 147–164, 1980.

[7] ____, "An optimal autoregressive spectral estimate," *Ann. Statist.*, vol. 9, pp. 300–306, 1981.

[8] ____, "Consistency of model selection and parameter estimation," *Essays in Time Series and Allied Processes*, J. Gani and M. B. Priestly, Eds. *J. Appl. Prob. Spec.*, vol. 23A, pp. 127–141, 1986.

[9] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[10] ____, "Consistent order estimates of autoregressive processes by shortest description of data," in *Analysis and Optimization of Stochastic Systems*. New York: Academic, 1980.

[11] ____, "Estimation of structure by minimum description of length," *Circuits, Syst., Signal Processing*, vol. 1, no. 3–4, pp. 395–406, 1982.

[12] ____, "Stochastic complexity and modeling," *Ann. Stat.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[13] E. Parzen, "Some recent advances in time series modeling," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 723–730, 1974.

[14] E. J. Hannan, "The estimation of the order of an ARMA process," *Ann. Statist.*, vol. 8, pp. 1071–1081, 1980.

[15] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc., Ser. B*, 41, no. 2, pp. 190–195, 1979.

[16] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[17] H. Tong, "Determination of the order of a Markov chain by Akaike's information criterion," *J. Appl. Prob.*, vol. 12, pp. 488–497, 1975.
[18] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, 1985.
[19] P. M. T. Broersen, "Selecting the order of autoregressive models from small samples," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-33, no. 4, pp. 874–879, 1986.
[20] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," Technion—Israel Inst. Technol., Haifa, Tech. Rep. EE 648, Nov. 1987, also to appear in *IEEE Trans. Inform. Theory*, vol. IT-35, Sept. 1989.
[21] J. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[22] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.
[23] B. Efron and D. Traux, "Large deviations theory in exponential families," *Ann. Math. Statist.*, vol. 39, no. 5, pp. 1402–1424, 1968.
[24] M. Gutman, "On tests for randomness, tests for independence, and universal data compression," submitted to the *IEEE Trans. Inform. Theory*.
[25] ____, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 2, pp. 401–408, Mar. 1989.

Quantum Communication with Coherent States

C. BENDJABALLAH AND M. CHARBIT

***Abstract*—Quantum communication is studied under the cutoff rate criterion for transmission of M coherent states of real amplitude ($M = 2,\cdots,7$). Comparison with other measurement operators of interest, such as the number and the quasi-classical operators, is made. The effect of quantization of the decision level on the cutoff rate is also considered.**

I. Introduction

Recent technical advances in stable monomode light sources and photon detectors, e.g., semiconductor lasers and avalanche photodiodes, have increased interest in the study of systems for the transmission of information using photon counting techniques. It has been shown that these techniques perform better than the usual ones [1], particularly in applications to optical space [2]. This has stimulated important work among communication and information researchers to establish analytical results that describe the best mode of transmission of the message, and also to calculate the fundamental limits on the performance of such systems.

These limits are also important for some aspects of the theory because of various links between the quantum theory of information and the theory of measurement in quantum mechanics. For example, the generalization of the uncertainty relations via quantum entropy [3], [4] is of great importance in both domains. On the other hand, the experimental interest of these fundamental limits is also evident since, when building practical receivers, a comparison of the performance attained by the device in operation with that of the ideal system is necessary. For this purpose one of the criteria widely used to characterize the performance is the channel capacity [5].

Although important progress in the mathematical theory of the quantum communication channel has been made in recent years, mainly through the work of Davies [6], Holevo [7], [8], and Helstrom [9], there are still several questions to be studied. One of these is the calculation of the quantum channel capacity from the mutual information

$$I\{P_{ij}\} = \sum_{ij} E(P_{ij}) - \sum_{i=1}^{M} E\left(\sum_{j=1}^{M} P_{ij}\right) - \sum_{j=1}^{N} E\left(\sum_{i=1}^{M} P_{ij}\right). \quad (1.1)$$

The function $E(P_{ij})$ is the usual entropy

$$E(P_{ij}) = -P_{ij}\ln(P_{ij}) \quad (1.2)$$

with

$$P_{ij} = \Pr(\text{output digit } j | \text{digit } i \text{ sent}) = q_i \operatorname{tr}(\boldsymbol{O}_i \boldsymbol{R}_j) = q_i p_{ij} \quad (1.3)$$

and where p_{ij} is the transition probability between input density operator ($\boldsymbol{O}_i$) of prior probability q_i and output measurement operator ($\boldsymbol{R}_j$).

Applying Shannon's relation directly, we define the channel capacity as

$$C_0 = \max_{(\{q_i\},\{\boldsymbol{R}_j\},N)} I(\{\boldsymbol{O}_i\};\{\boldsymbol{R}_j\}). \quad (1.4)$$

Note that, because of the nonlinearity of the logarithmic function appearing in (1.2), the methods used in quantum detection theory [9] are not applicable to maximizing $I\{(\boldsymbol{O}_i),(\boldsymbol{R}_j)\}$. Some rigorous results of general interest have been established but concern only a few special situations [10]–[14]. Nevertheless, the channel capacity is not the only criterion of performance, and as already observed by Wozencraft and Jacobs [15] and later by Massey [16], the cutoff rate is also meaningful. This parameter determines a range of rates where reliable transmission of information is possible and also provides an insight into the modulation complexity when there is an error probability. In this sense the cutoff rate criterion is more informative than the channel capacity. The purpose of this correspondence is to derive some new results in the quantum theory of the cutoff rate.

In Section II basic relations of the usual equations of the optimal detection operator and the cutoff rate are recalled. Quantum definitions of these quantities are given, and the problem of optimizing them is formulated. Such optimization requires, as usual, constraints on the prior probabilities of the input states. Also, although algebraic calculations are straightforward, they become involved for $M > 5$ so that only a limited input alphabet can be simply analyzed. There are several particular channels for which a complete study is possible for any value of M. One of these is considered in Appendix III. Approximate expressions for large M are also available. However, since the optimization is often difficult, it is shown in Section II that upper bounds are very useful. These bounds are found in Section III for input coherent states and are briefly compared in Section IV with those derived from semiclassical operators [15]–[17], [23], [24].

Manuscript received March 19, 1987; revised August 1988. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Ann Arbor, MI, October 6–9, 1986, in part at COCT'87, Karuizawa, Japan, August 24–27, 1987, and in part at the IEEE International Symposium on Information Theory, Kobe, Japan, June 19–24, 1988. This work was supported by E.N.S. Telecommunications (U.A. 820 du C.N.R.S.), Paris, France.

The authors are with the Laboratoire des Signaux et Systèmes du C.N.R.S., Ecole Supérieure d'Electricité, Plateau de Moulon, 91192, Gif sur Yvette Cedex, France.

IEEE Log Number 8930783.

0018-9448/89/0900-1114$01.00 ©1989 IEEE