U·M·I Dissertation
Information Service

University Microfilms International
A Bell & Howell Information Company
300 N. Zeeb Road, Ann Arbor, Michigan 48106

Kumar, Anand Ramachandran

A DISTRIBUTION-FREE MODEL ORDER ESTIMATION TECHNIQUE USING ENTROPY

Rice University M.S. 1986

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

RICE UNIVERSITY

A DISTRIBUTION-FREE MODEL ORDER ESTIMATION TECHNIQUE USING ENTROPY

by

ANAND RAMACHANDRAN KUMAR

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

APPROVED, THESIS COMMITTEE:

Don H. Johnson

Don H. Johnson, Associate Professor
Department of Electrical & Computer Engineering
Director

Behnam Azhang

Behnaam Aazhang, Assistant Professor
Department of Electrical & Computer Engineering

George R. Terrell, Lecturer
Department of Mathematical Sciences

Houston, Texas

December, 1985

A Distribution-Free Model Order Estimation Technique using Entropy

by

Anand Ramachandran Kumar

Abstract

A new model order determination procedure using concepts of entropy is proposed; this procedure (Entropy Method) makes few assumptions on the model and the allowable class of inputs. Simulations were performed to determine the performance characteristics of the Entropy Method on first order Autoregressive (AR) models (Gaussian and Exponential) and on a first order nonlinear model. The Entropy Method performed well on Gaussian AR time series for large values of first order coefficients; it performed very well on Exponential AR time series for most choices of coefficients. It was further observed that this technique estimated the model order of data with nonlinear dependence structure. The performace analysis revealed three limitations: the method is not sensitive to low dependence, the technique requires a large sample size to estimate the model order and the computational resources required are enormous.

Acknowledgements

I would like to express my gratitude to Dr. Don Johnson for his contribution to this thesis and his contribution to the growth in my research capabilities. I also appreciate his constant supply of morale boosters. I would also like to thank George Terrell for providing the insight that set the ball rolling.

I would like to thank at least a few friends who have contributed to this thesis in different ways. The algorithm for a key program evolved out of discussions with Darel Linebarger and Mike Heideman. Mike's word processing expertise was a source that I constantly tapped. Discussions with Doug Jones, especially in the initial stages helped a great deal in clearing the air.

Doug Jones and Henrik Sorensen are friends to have around, especially when things are not going well. On the home front I was and am fortunate to have such wonderful friends like Senthil and Jeyak who displayed great restraint in putting up with my, often, abominable behavior; they also provided the constant encouragement that helped me to hang in there.

I wish to acknowledge the NIH grant that funded my research.

Table of Contents

CHAPTER 1

Introduction

The model order is a measure of the complexity of the model, implying how many parameters need to be estimated. Thus determinaton of the model order is the key to statistical model identification. Model identification in turn is the first step to system identification. Linear, time-invariant models are most commonly assumed; from a practical point of view this assumption considerably reduces the complexity of the identification problem. The Autoregressive (AR) model and the Moving Average (MA) model are two commonly assumed linear models. The Minimum AIC Estimate (MAICE) technique proposed by Akaike [1] and the Minimum Description Length (MDL) criterion derived by Schwarz [17] are two well known model order determination procedures; both of these techniques assume a linear model and the methods fail, as will be shown, when this assumption is not true. Further analytical results for the two methods are possible only when the time series elements are jointly normal. But linear models (restricted to Gaussian inputs) are not appropriate in certain situations; for example, in the analysis of spike trains from the Lateral Superior Olive (LSO) of a cat nonlinear models with exponential inputs are appropriate [8].

This work proposes a new procedure - that makes few assumptions on the model and the input - to determine the model order of a given sequence of observables (i.e., a time series) using the concept of entropy.

CHAPTER 2

Background Material

2.1. Model Order Determination

Let $\{Y_n\}$ be a stochastic time series. We assume that this time series is derived from the time series $\{X_n\}$ with a transformation having the Markovian form

$$Y_n = f(Y_{n-1}, Y_{n-2}, \ldots, Y_{n-p}, X_n, \theta).$$

The input sequence $\{X_n\}$ is assumed to consist of statistically independent, identically distributed random variables. Clearly, under the above transformation Y_n does not depend on Y_i for all $i < n-p$ and $i > n$. p is thus defined to be the *order* of this assumed model. $f(\cdot)$ is the assumed functional relationship between $\{X_n\}$ and $\{Y_n\}$. The components $\{\theta_{n,i}\}$, i=1, 2,..., k of the vector θ are defined to be parameters of the model. The parameters $\{\theta_{n,i}\}$ may be constants, implying a stationary model, or may vary with time (index n), which result in a non-stationary time series $\{Y_n\}$. The number of model parameters (k) need not equal the model order (p).

Perhaps the simplest example occurs when $f(\cdot)$ has a multi-linear form

$$Y_n = \theta_1 Y_{n-1} + \theta_2 Y_{n-2} + \ldots. + \theta_p Y_{n-p} + X_n.$$

In the statistical literature, this model is termed an autoregressive (AR) model [4] and in the signal processing literature an all pole model [10]. The moving average (MA) and the autoregressive moving average (ARMA) models are other commonly used linear models [4]. In the AR model the number of model parameters equals the model order.

Usually, the observables consist only of the output time series $\{Y_n\}$. If the characteristics of the input $\{X_n\}$ (i.e., its amplitude distribution) are known, the signal processing problem is to determine the characteristics of the transformation $f(\cdot)$. For example, under the linear assumption, if p is known, the determination of the parameter vector θ is an important problem in speech signal processing. Another example is to make few assumptions on $f(\cdot)$ - define a membership class - and determine the model order p and then the parameter vector θ. This latter problem is of concern here; in particular, a linear form for $f(\cdot)$ will *not* be assumed.

The determination of p is dependent on the joint probability distribution of the observed time series $\{Y_n\}$. If $f(\cdot)$ is multi-linear and $\{X_n\}$ is normal then $\{Y_n\}$ is normal. This example is one of the few for which the joint distribution of $\{Y_n\}$ is known. Even if $f(\cdot)$ is linear, determination of this joint distribution when $\{X_n\}$ is non-Gaussian is difficult. When $f(\cdot)$ is nonlinear, the analytic situation is worse. There are very few nonlinear models used because analysis is extremely difficult, if not impossible. The following input-output relationship is an example of a nonlinear model that is of importance here.

$$Y_n = \theta_1 e^{-\theta_2 Y_{n-1}} + \theta_3 e^{-\theta_4 Y_{n-2}} + \cdots + \theta_{2p-1} e^{-\theta_{2p} Y_{n-p}} + X_n \tag{2.1}$$

Whether linear or nonlinear models are to be studied, the model order determination problem - given observation of the sequence $\{Y_n\}$, determine the value of p - is the key to the identification of the model. The model order is a measure of the complexity of the model, implying how many parameters need to be estimated and defining the dimension of the joint distribution of $\{Y_n\}$ necessary to capture its statistics. The well-known model order determination procedures essentially assume a linear model. Primary among them are the MAICE and MDL procedures.

Minimum AIC Estimate (MAICE):

The following information criterion was defined by Akaike [1]

$$\mathrm{AIC}(\hat{\theta}) = (-2)\log(\text{maximum likelihood}) + 2k.$$

The above criterion contains an estimate of minus twice the expected log likelihood of the model whose parameters are determined by the method of maximum likelihood. The natural logarithm is implied in the above definition. k is the number of parameters of the model that are independently adjusted to obtain the estimate $\hat{\theta}$. In the context of model order determination the estimate $\hat{\theta}$ is the classical maximum likelihood estimate obtained by maximizing the conditional density function $p_{Y_1,\ldots,Y_n|\theta}(y_1,\ldots,y_n|\theta)$ over all vectors θ. The first term in the above criterion decreases with k and the second term increases. Therefore a minimum will occur. The value of k that best fits the data results in the minimum value for AIC; this value will be selected as the model order that best fits the data. As remarked earlier the number of model parameters (k), in general, is not equal to the model order (p). The only way this technique can be used as a model order determining method is to assume $k=p$. Clearly the method will fail when this assumption is not true.

As mentioned earlier, rarely is the joint distribution of $\{Y_n\}$, and hence the likelihood function, known. In the linear, Gaussian case the equivalent quantity to be minimized in a p^{th} order AR model is [10]

$$\log(V_i) + \frac{2i}{N_e}.$$

where V_i is the i^{th} normalized prediction error and N_e is the effective number of data points used. The effective width of the analysis window can be taken as the ratio of the energy under that window relative to that under a rectangular window. For a Hanning window (the window used in subsequent analysis), $N_e = 0.374N$ (N is the actual number of data points).

Minimum Description Length (MDL):

This criterion [17] chooses the model which minimizes

$$(-2)\ \log(\text{maximum likelihood}) + k\log N.$$

In the above expression, N is the sample size and k denotes the number of parameters of the model independently adjusted in obtaining the maximum likelihood estimate (as in MAICE). The criterion is a Bayesian solution to the problem of selecting the model that is *a posteriori* most probable. The observations are assumed to come from a Koopman-Darmois family which essentially restricts the class of observations considered to those which arise from a linear model.

Performance of MAICE and MDL:

To assess the characteristics of these model order determination procedures on data for which they are intended (linear model with Gaussian input) and for which they have little justification (linear model with non-Gaussian input and nonlinear model with non-Gaussian input), a series of simulations were performed on first-order models. The performance measure used was the model order selected in repeated simulations (runs). The results of simulations with a first order Gaussian AR model (Table 2.1) indicate that the MAICE technique, for a first order coefficient larger in magnitude than 0.1, chooses the correct order of the model in about 80 percent to 86 percent of the runs; the corresponding simulations for MDL indicate that the correct model order was selected in greater than 99 percent of the runs. It is evident from the average value and the standard deviation of the model order chosen that the MAICE technique tends to overestimate the model order.

Table 2.1
Performance of MAICE and MDL on First Order AR Time Series

ρ	Percentage of Correct Order Prediction			
	Gaussian		Exponential	
	MAICE	MDL	MAICE	MDL
-0.1	85 (1.30, 1.01)	97 (0.99, 0.17)	84 (1.23, 0.60)	100 (1.00, 0.00)
-0.2	82 (1.34, 0.92)	100 (1.00, 0.00)	85 (1.28, 0.74)	100 (1.00, 0.00)
-0.3	85 (1.32, 0.98)	100 (1.00, 0.00)	90 (1.24, 0.98)	100 (1.00, 0.00)
-0.4	83 (1.31, 0.81)	100 (1.00, 0.00)	93 (1.09, 0.35)	100 (1.00, 0.00)
-0.5	81 (1.33, 0.79)	100 (1.00, 0.00)	88 (1.25, 0.78)	100 (1.00, 0.00)
-0.6	80 (1.44, 1.12)	100 (1.00, 0.00)	89 (1.17, 0.53)	100 (1.00, 0.00)
-0.7	84 (1.33, 0.95)	100 (1.00, 0.00)	85 (1.20, 0.53)	100 (1.00, 0.00)
-0.8	85 (1.27, 0.87)	100 (1.00, 0.00)	86 (1.25, 0.71)	100 (1.00, 0.00)
-0.9	86 (1.27, 0.77)	99 (1.01, 0.10)	88 (1.25, 0.85)	100 (1.00, 0.00)

(ρ ($\equiv \theta_1$) is the coefficient of the first order AR time series. N = 6000. Percentages based on 100 simulations. The numbers in brackets are the mean and standard deviation of the model order chosen in the 100 runs)

For a non-Gaussian autoregressive time series there exists no framework to obtain an explicit mathematical expression for the log likelihood function; one *adhoc* approach would be to apply the Gaussian analysis to the non-Gaussian AR time series. The performance of MAICE on a first order exponential autoregressive time series (Table 2.1) is marginally superior to the Gaussian case for most choices of first order coefficients. It is evident from the average value and the standard deviation of the model order chosen that the MAICE technique tends to overestimate the model order. The corresponding results in Table 2.1 indicate that the MDL technique performs as well on the exponential AR time series as on the Gaussian AR time series.

It is not clear how one could use these methods on data with a nonlinear dependence structure. One approach would be to assume a Gaussian AR model and apply the above technique. A first order time series with an exponential (nonlinear) dependence structure (see equation (2.1)) was generated by the technique discussed by Johnson and Linebarger [7]. Independent, exponentially distributed random variables were used as the input to generate the nonlinearly dependent time series. In the simulations p was unity and one of the parameters of the model ($\theta_{1,2}$) was held constant, while $\theta_{1,1}$ was allowed to vary; in this manner the effect of increasing nonlinear dependence on the performance of the techniques under analysis is more easily seen (for then the dependence is directly related to $\theta_{1,1}$). The results of the simulations have been tabulated in Table 2.2. The serial correlation coefficient, ρ (which is the first order coefficient under the assumption of an AR model) for was computed for the data and has been included in the table. The results of the simulations indicate that MAICE and MDL do very poorly even on data with moderately large nonlinear dependence.

Table 2.2
Performance of MAICE and MDL on a Time Series with
First Order Nonlinear Dependence

$\theta_{1,1}$	ρ	Percentage of Correct Order Prediction	
		MAICE	MDL
.002	-0.12	81 (1.32, 0.79)	100 (1.00, 0.00)
.003	-0.17	73 (1.43, 1.01)	98 (1.02, 0.14)
.006	-0.30	7 (2.29, 1.17)	70 (1.30, 0.46)
.009	-0.40	0 (2.22, 0.76)	5 (1.95, 0.22)
.012	-0.51	0 (2.33, 0.62)	0 (2.00, 0.00)
.015	-0.59	0 (2.92, 1.14)	0 (2.03, 0.22)
.02	-0.69	0 (3.59, 1.02)	0 (2.31, 0.46)

(ρ is the serial correlation coefficient. $\theta_{1,2}$ was
held constant at 195.4. N = 6000. Percentages based on 100 simulations.
The numbers in brackets are the mean and standard deviation of the
model order chosen in the 100 runs)

It is evident that the two model order determining techniques studied thus far - MAICE and MDL - are appropriate only for data with linear dependence and are not suitable for nonlinearly dependent data. The dependence of these techniques on the prediction error explain their poor performance on nonlinearly dependent data: the prediction error is completely determined by the correlation matrix. Only in the case when $\{Y_n\}$ are jointly Gaussian is the time series completely described by the first and second moments; in this restricted case, lack of linear correlation implies statistical independence. There can be statistical dependence in a time series that is not jointly Gaussian even in the absence of linear correlation. Thus, correlation is not a good measure of nonlinear dependence; this results in the methods choosing incorrect model orders. One other major drawback is that they can be used as model order determining techniques only if the number of model parameters (k) is related to the model order (p) in a known manner.

2.2. Measures of Dependence

Quantities that characterize the strength of dependence between two random variables by a numerical value are termed as measures of dependence. The applicability of these measures to model order determination will be investigated in this section. The time series $\{Y_n\}$ has, by definition, p^{th} order Markovian dependence if and only if

$$\begin{aligned} &p_{Y_i \mid Y_{i-1}, Y_{i-2}, \ldots, Y_1}(y_i \mid y_{i-1}, y_{i-2}, \ldots, y_1) \\ &\qquad = p_{Y_i \mid Y_{i-1}, Y_{i-2}, \ldots, Y_{i-p}}(y_i \mid y_{i-1}, y_{i-2}, \ldots, y_{i-p}). \end{aligned} \tag{2.2}$$

Using Bayes rule one also obtains

$$p_{Y_i \mid Y_{i-1}, \ldots, Y_{i-p-1}}(y_i \mid y_{i-1}, \ldots, y_{i-p-1}) = \frac{p_{Y_i, Y_{i-p-1} \mid Y_{i-1}, \ldots, Y_{i-p}}(y_i, y_{i-p-1} \mid y_{i-1}, \ldots, y_{i-p})}{p_{Y_{i-p-1} \mid Y_{i-1}, \ldots, Y_{i-p}}(y_{i-p-1} \mid y_{i-1}, \ldots, y_{i-p})}.$$

Thus for a time series with p^{th} order Markovian dependence we have

$$\begin{aligned} &p_{Y_i, Y_{i-p-1} \mid Y_{i-1}, \ldots, Y_{i-p}}(y_i, y_{i-p-1} \mid y_{i-1}, \ldots, y_{i-p}) \\ &\qquad = p_{Y_i \mid Y_{i-1}, \ldots, Y_{i-p}}(y_i \mid y_{i-1}, \ldots, y_{i-p})\, p_{Y_{i-p-1} \mid Y_{i-1}, \ldots, Y_{i-p}}(y_{i-p-1} \mid y_{i-1}, \ldots, y_{i-p}). \end{aligned}$$

Thus in model order determination one is essentially testing for independence of Y_i and Y_{i-p-1} conditioned on $\{Y_{i-1}, \ldots, Y_{i-p}\}$. Thus to determine model order one should check for the independence of Y_i and Y_{i-l-1} conditioned on $\{Y_{i-1}, \ldots, Y_{i-l}\}$ for increasing values of l; the value of l that results in independence is the model order.

We shall consider four measures of dependence between two random variables, X and Y, - the correlation coefficient, correlation ratios, maximal correlation and mean square contingency. Renyi [15] laid down seven postulates that a dependence measure should satisfy; they are (the general dependence measure will be denoted by $\delta(X, Y)$)

(i) $\delta(X, Y)$ is defined for any pair of random variables neither of which are constant with probability one.

(ii) $\delta(X, Y) = \delta(Y, X)$.

(iii) $0 \leq \delta(X, Y) \leq 1$.

(iv) $\delta(X, Y) = 0$, if and only if X and Y are independent.

(v) $\delta(X, Y) = 1$, if there is a strict dependence between X and Y, i.e., either $X = g(Y)$ or $Y = f(X)$ where $g(\cdot)$ and $f(\cdot)$ are Borel-measurable functions.

(vi) If the Borel-measurable functions $f(\cdot)$ and $g(\cdot)$ map $\Re^1$ in a one-to-one way onto itself, $\delta(f(X), g(Y)) = \delta(X, Y)$.

(vii) If the joint distribution of X and Y is normal, then $\delta(X, Y) = |\rho(X, Y)|$ where $\rho(X, Y)$ is the correlation coefficient of X and Y.

(viii) In addition to the above seven postulates we require that the dependence measure be computable for a given time series.

It should be noted that of the above mentioned properties, in the context of model order determination, only (iv) and (viii) are essential while all others are desirable. The range of $\delta(X, Y)$ need only be finite (not necessarily [0,1] as in (iii)).

The correlation coefficient, $\rho(X, Y)$ is defined by [15]

$$\rho(X, Y) = \frac{E(XY) - E(X)\,E(Y)}{\sqrt{Var(X)\,Var(Y)}}$$

provided $Var(X)$ and $Var(Y)$ are finite and nonzero; $E(\cdot)$ denotes the expectation and $Var(\cdot)$ is the variance. This dependence measure satisfies only properties (ii), (iii), (vii) and (viii) in the above list. This quantity is an inadequate measure of nonlinear dependence; for example suppose X is uniformly distributed on $(-1, 1)$ and $Y = 5X^3 - 3X$ then $\rho(X, Y) = 0$, implying independence (i.e., it does not satisfy (iv)).

The correlation ratio is defined by [15]

$$\Theta(X, Y) = \max \left[\Theta_X(Y), \Theta_Y(X)\right]$$

where

$$\Theta_X(Y) = \frac{\sqrt{Var(E(Y|X))}}{\sqrt{Var(Y)}}$$

provided $Var(Y)$ exists and is nonzero. $E(\cdot|\cdot)$ denotes conditional expectation. It does not satisfy properties (i), (vii) and more importantly (iv); $\Theta(X, Y)$ is zero when (X, Y) is uniformly distributed in a circle and hence suffers the same failing as the correlation coefficient, $\rho(X, Y)$, i.e., it is not an adequate measure of general dependence.

The maximal correlation, $S(X, Y)$ is defined to be [15]

$$S(X, Y) = \sup_{f, g} \rho(f(X), g(Y))$$

where $f(\cdot)$ and $g(\cdot)$ run over all Borel-measurable functions such that $\rho(f(X), g(Y))$ makes sense, i.e., $f(X)$ and $g(Y)$ have finite and nonzero variance. The maximal correlation is superior, mathematically, to the other dependence measures (it satifies postulates (i) through (vii)); however, it is diffcult to obtain a mathematical expression for $S(X, Y)$ even when one knows the joint distribution of the two random variables. The analytical situation is worse when the joint distribution is not known. Therefore estimating this quantity for two random variables from a time series will be difficult if not impossible.

Then mean square contingency of X and Y, with continuous joint distribution, is defined to be [14]

$$C(X, Y) = \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(p(x, y) - p(x)\,p(y))^2}{p(x)\,p(y)}\, dx\, dy\right]^{\frac{1}{2}}.$$

The following quantity

$$\frac{C(X,Y)}{\sqrt{1+C(X,Y)^2}}$$

is a dependence measure that satisfies all properties except (i) and (v); it is clear that this dependence measure involves estimating probability density functions and, as will be discussed in *§* 1.3, this is not desirable. Hence this measure does not satisfy property (viii).

The correlation coefficient and correlation ratio measures are not appropriate for we are interested, as is evident from our discussions in § 2.1, in determining the model order of a time series with nonlinear dependence. The two dependence measures that are theoretically sound, maximal correlation and mean square contingency, are not computationally feasible and hence are not appropriate candidates for determining model order; the computational complexity is increased by the fact that in determining model order we are testing for a *conditional* independence between Y_i and Y_{i-l-1} (discussed earlier).

2.3. Information Theory: Basics and Application to Model Order Determination

The discussion in § 2.1 suggests that MAICE and MDL are not adequate tools to determine Markovian order when nonlinear dependence is present. The discussion in the previous section indicates that the four measures of dependence - correlation coefficients, correlation ratios, maximal correlation and mean square contingency - are not appropriate either. An informational measure statistic of dependence was discussed by E. H. Linfoot [9]; this approach suggests a information-theoretic basis for determining the Markovian order of dependence in a time series. Nakahama et al. used the concept of entropy to determine Markovian order [11].

Let X be a discrete random variable, taking on the distinct values x_k with the probabilities p_k (k = 1,2,...), i.e.

$$Pr(X{=}x_k) = p_k \text{ where } p_k \geq 0 \text{ and } \sum_{k=1}^{\infty} p_k = 1.$$

Here $Pr(\cdot)$ denotes the probability of the event in the brackets. The entropy of X (which may also called the entropy of the probability distribution of X) as defined by Shannon [18] is denoted to be

$$H(X) = -\sum_{k=1}^{\infty} p_k \log p_k,$$

provided the series on the right of the equation converges. When X has a continuous distribution function (with a probability density function $p(x)$), the entropy is defined to be

$$H(X) = -\int p(x) \log p(x)\, dx, \tag{2.3}$$

provided the integral has a finite value; the integration is performed over the region for which the probability density function is defined. Natural logarithm is implied in the above definition. To the knowledge of the author there exists no continuous distribution, for which the above integral is not finite. But in practice, the time series recorded may have a distribution that is not truly continuous (i.e., it may have impulses), although the stochastic process generating the time series may have a continuous distribution and thus the above integral may not be finite for the distribution associated with the time series. This definition is easily extended to a random vector

$$H(X) = -\int p(x) \log p(x)\, dx \tag{2.4}$$

where $X \in \Re^d$ and the integration is performed over $\Re^d$. The above definition can also be viewed as the joint entropy of the components of the random vector (i.e. random variables).

H(X) has the following properties [20]:

(i) H(X) can be positive, negative or zero.

For a Gaussian random vector X one has

$$H(X) = \log \left[(2\pi e)^{\frac{d}{2}} \; |K|^{\frac{1}{2}} \right] \tag{2.5}$$

where K is the covariance matrix and $|\cdot|$ denotes the determinant. It is quite clear that depending on the value of the determinant of the covariance matrix H(X) could be positive, negative or zero.

(ii) H(X) depends on the coordinate system.

Let H(X) denote the entropy of the random vector X. Suppose Y is the vector obtained by an invertible transformation of the vector X. The entropy of Y is given by

$$H(Y) = H(X) - E_X\{\log |J(X/Y)|\}.$$

$J(X/Y)$ is the Jacobian of the transformation from X to Y and $E_X\{\cdot\}$ denotes the expectation operation with respect to the distribution of X.

(iii) H(X) is invariant to translation.

The Jacobian for an invertible translation transformation is unity and hence the entropy of a random vector is translation invariant.

The joint entropy of two random vectors, X and Y is defined to be

$$H(X, Y) = -\int p(x, y) \log (p(x, y))dxdy.$$

where $p(\cdot)$ is the joint distribution of the random vectors X and Y. Conditional entropy is defined to be

$$H(X \mid Y) = -\int p(x, y) \log (p(x \mid y))dxdy. \tag{2.6}$$

where $p(\cdot \mid \cdot)$ is the conditional distribution of X given Y. The conditional entropy is related to the joint entropy by

$$H(X \mid Y) = H(X, Y) - H(Y). \tag{2.7}$$

The use of conditional entropy in model order determination depends on the theorem that the conditional entropy equals $H(X)$, if and only if the random vectors X and Y are statistically independent. First suppose $H(X \mid Y) = H(X)$. Thus,

$$-\int p(x, y) \log p(x \mid y)dxdy = -\int p(x) \log p(x)dx.$$

Using the definition of conditional densities, we have

$$-\int p(x, y) \log p(x, y)dxdy + \int p(y) \log p(y)dy = -\int p(x) \log p(x)dx.$$

Combining terms one obtains

$$\log \left[\frac{p(x, y)}{p(x)p(y)} \right] = 0 \textit{ almost everywhere (a.e.)}$$

implying $p(x, y) = p(x)p(y)$ *a.e.*, which means that X and Y are statistically independent. Conversely, suppose that X and Y are statistically independent; then by definition $p(x \mid y) = p(x)$. Substituting into equation (2.6), one easily obtains $H(X \mid Y) = H(X)$. Thus when X and Y are statistically independent, equation (2.7) reduces to

$$H(X, Y) = H(X) + H(Y). \tag{2.8}$$

The concept of conditional entropy can be used to determine the model order of a time series $\{Y_1, Y_2, \ldots, Y_n\}$. Applying the definition of Markov property of a time series (equation (2.2)) to the expression for conditional entropy in equation (2.6) one obtains the following equivalent definition of p^{th} order Markovian dependence:

$$H(Y_i \mid Y_{i-1}, Y_{i-2}, \ldots, Y_1) = H(Y_i \mid Y_{i-1}, \ldots, Y_{i-p}),$$

if and only if the model order is p. Thus for a time series with p^{th} order Markovian dependence one has the following relationship

$$H(Y_i \mid Y_{i-1}, \ldots, Y_{i-p}) = H(Y_i \mid Y_{i-1}, \cdots, Y_{i-p-l}) \text{ for every } l \geq 1. \tag{2.9}$$

Now consider the difference $H(Y_i \mid Y_{i-1}, \ldots, Y_{i-l}) - H(Y_i \mid Y_{i-1}, \ldots, Y_{i-m})$ for $m \leq l$. From the definition of conditional entropy (equation (2.6)) this difference equals

$$\int p(Y_i, \ldots, Y_{i-l}) \log \frac{p(Y_i \mid Y_{i-1}, \ldots, Y_{i-m})}{p(Y_i \mid Y_{i-1}, \ldots, Y_{i-l})} \, dY_i \cdots dY_{i-l}.$$

After some manipulations and using Jensen's inequality [16] one obtains

$$\begin{aligned} &H(Y_i \mid Y_{i-1}, \ldots, Y_{i-l}) - H(Y_i \mid Y_{i-1}, \ldots, Y_{i-m}) \\ &\qquad\qquad \leq \log \int p(Y_i \mid Y_{i-1}, \ldots, Y_{i-m}) \, p(Y_{i-1}, \ldots, Y_{i-l}) \, dY_i \cdots dY_{i-l} \leq 0. \end{aligned}$$

This establishes the following inequalities

$$H(Y_i) \geq H(Y_i \mid Y_{i-1}) \geq \ldots\ldots \geq H(Y_i \mid Y_{i-1}, \ldots, Y_{i-p}).$$

Thus following monotonic relationship holds for any time series $\{Y_n\}$ with a p^{th} order Markovian dependence

$$H(Y_i) \geq H(Y_i \mid Y_{i-1}) \geq \ldots\ldots \geq H(Y_i \mid Y_{i-1}, \ldots, Y_{i-p}) = H(Y_i \mid Y_{i-1}, \ldots, Y_{i-p-1}) = \cdots \tag{2.10}$$

Thus the model order for the given time series, $\{Y_n\}$, can be determined by computing the conditional entropy, $H(Y_i | Y_{i-1}, \ldots, Y_{i-m})$, for increasing values of m (starting with unity) and the value of m for which the difference $H(Y_i | Y_{i-1}, \ldots, Y_{i-l}) - H(Y_i | Y_{i-1}, \ldots, Y_{i-l-1})$ is zero for all $l \geq m$ is the model order. Note that one has to compute up to a $(p+1)^{th}$ order conditional entropy to estimate the model order of a p^{th}-order time series.

We will now digress a little to view entropy as a measure of dependence in the framework of the eight postulates discussed in § 2.2. Consider the quantity $H(Y|X) - H(Y)$ as a measure of dependence between the random variables X and Y. This quantity satisfies the postulates (i), (ii), (iv) and (vi). It does not satisfy property (iii) but the quantity has a finite range and as mentioned in § 2.2 this will suffice in the model order determination context. As will be seen in *§* 3.1 this quantity can be computed without estimating explicitly the joint probability density. Thus the crucial postulates (iv) and (viii) are satisfied. These two crucial postulates hold for the quantity $H(Y|X) - H(X)$ (for any vector X) as well and hence this information-theoretic quantity, as has already been suggested, can be used to estimate the model order of a time series.

To use the conditional entropy, one must be concerned with the computational details. One obvious approach is to estimate explicitly the probability density function, compute the joint entropy, and obtain the conditional entropy from equation (2.7). Nakahama et al. defined the m^{th} order dependency D_m of the time series [11] $\{Y_i\}$ as

$$D_m = \frac{H(Y_i) - H(Y_i | Y_{i-1}, Y_{i-2}, \ldots Y_{i-m})}{H(Y_i)}. \tag{2.11}$$

For all m, D_m takes on values in the interval [0,1]; this easily follows from the inequalities of equation (2.10). From equations (2.10) and (2.11) the following inequality for D_m is obtained

$$D_0 \leq D_1 \leq \ldots\ldots \leq D_p = D_{p+1} = \ldots\ldots$$

Clearly, when there is no dependence in the time series, $H(Y_i) = H(Y_i | Y_{i-1}, \ldots, Y_{i-m})$ and hence D_m is zero for all values of m. When the statistic $\Delta D_m = D_m - D_{m-1}$ is zero for all $m \geq l+1$, the

D_m is zero for all values of m. When the statistic $\Delta D_m = D_m - D_{m-1}$ is zero for all $m \geq l{+}1$, the least value of l gives the model order, p.

To compute the entropy terms in equation (2.11), histogram estimates of the joint density, $\hat{p}(Y_i, \ldots, Y_{i-m})$ (for different values m) were computed. The binwidth was chosen in the following manner: $binwidth = \frac{6\,\sigma}{n_b}$, where σ was the standard deviation of the time series $\{Y_n\}$ and n_b was the number of bins used. The estimates of the corresponding joint entropies, $\hat{H}(Y_i, \ldots, Y_{i-m})$ (for different values of m) were then computed from the density estimates in the obvious way. The conditional entropy, $\hat{H}(Y_i \,|\, Y_{i-1}, Y_{i-2}, \ldots, Y_{i-m})$ (for different values of m) was then obtained from the joint entropies using equation (2.7). The estimate of the m^{th}-order dependency, $\hat{D}_m$, was then computed using the formula in equation (2.11).

To test the hypotheses $\Delta\hat{D}_m = 0$ and $\Delta\hat{D}_m \neq 0$, the sampling distribution of this statistic under the null hypothesis is required. Lacking an analytic expression for the general case, the authors of the paper obtained the distribution for this statistic empirically. They estimated the distribution for the statistic $\Delta\hat{D}_m^{sh} = \hat{D}_m^{sh} - D_{m-1}$, where $\hat{D}_m^{sh}$ was the estimate of the m^{th}-order dependency of the set $\{Y_i, \ldots, Y_{i-m+1}, Y_{i-m}^{sh}\}$, where Y_{i-m}^{sh} was drawn randomly from the original time series $\{Y_n\}$ (this procedure is termed as shuffling). The assumption was that $\{Y_{i-m}^{sh}\}$ and $\{Y_i, \ldots, Y_{i-m+1}\}$ are statistically independent and the expected value of $\hat{p}(Y_i \,|\, Y_{i-1}, \ldots, Y_{i-m+1}, Y_{i-m}^{sh})$ approximately equals $\hat{p}(Y_i \,|\, Y_{i-1}, \ldots, Y_{i-m+1})$. The shuffling was repeated a number of times and the distribution of $\Delta\hat{D}_m^{sh}$ was estimated for each m. The model order was estimated for three distinct choices of binwidths corresponding to $n_b = 2$, 3 and 5. The results of simulations performed by the authors indicate that, in some cases, the model order estimated varied with the binwidth.

In the model order determining method of Nakahama et al. the chosen binwidths were extremely crude; they used *at most* 5 bins on each axis to obtain the histogram estimates. The first

25 to estimate the second order density estimate, $p(Y_i, Y_{i-1})$ - were used in estimating the probability density function. A second comment is that there were very few sample points in each bin, particularly in the estimation of higher order probability density functions. To compute the model order for a time series with a second order dependence one needs to compute D_m for $m = 1, 2, 3$ and D_3 requires the computation of a fourth order probability density function, i.e., $p(Y_i, Y_{i-1}, Y_{i-2}, Y_{i-3})$. As in one of their more plentiful examples, if 4000 time series points were used to estimate the fourth order probability density function with 5 bins along each axis; this implies that there were, on an average, just $\frac{4000}{5^4} \cong 7$ points in each bin. The resulting heavily oversmoothed density estimates were used in estimating the dependency and hence the model order. Further it is not clear that the shuffling methods used in determining the null distribution were very reliable. In addition to these limitations the dependency method did not have a firm theoretical foundation. The paper did not include any performance analysis; for example, how often did the method estimate the model order correctly in repeated simulations. Another aspect of the procedure not covered in the paper is that of robustness, i.e., in what range of dependence levels did the method estimate the model order efficiently. These criteria must be used to assess the validity of any model order determining technique.

CHAPTER 3

Estimating Model Order Using Entropy

3.1. Estimating Entropy

It is clear from Chapter 2 that information theory methods can be used in model order determination. But the discussions on the dependency approach to model order determination suggest that an explicit estimation of the probability density function is not desirable. Does a procedure exist that estimates the joint entropy and hence the conditional entropy without explicitly estimating the probability density function? To ensure that the procedure is applicable to a wide variety of problems we also require that the statistic estimating the joint entropy not require any knowledge, at least asymptotically, of the underlying distribution of the given time series (i.e., a distribution-free statistic).

Suppose $\{Y_n, n = 1,..., N\}$ is the given observed time series (of identically distributed random variables). Let Y_n have a continuous amplitude distribution function $P_Y(\cdot)$ and associated probability density function $p_Y(\cdot)$. From elementary probability theory we have

$$\int_{Y_n}^{TY_n} p_Y(y)\, dy = P_Y(TY_n) - P_Y(Y_n) \quad . \qquad (3.1)$$

where T is an operator whose domain and range comprises of elements of the time series; it maps Y_n to the smallest (in amplitude) random variable larger than Y_n - TY_n - and is mathematically defined as

$$TY_n \in \{Y_n\} \text{ and } TY_n = arg\left\{\min_{Y_i > Y_n} (Y_i - Y_n)\right\}.$$

The following approximation to equation (3.1)

$$p_Y(Y_n)(TY_n - Y_n) \approx P_Y(TY_n) - P_Y(Y_n)$$

holds when the difference $TY_n - Y_n$ is small. When the empirical distribution function is substituted into the right hand side, we obtain

$$p_Y(Y_n)\,(TY_n - Y_n) \approx \frac{1}{N}. \tag{3.2}$$

The entropy of $p_Y(\cdot)$, given by equation (2.3), can be expressed as $E\left[-\log p_Y(\cdot)\right]$; the entropy of $p_Y(\cdot)$ can be estimated by the average value of $-\log p_Y(\cdot)$. Using equation (3.2) the following estimate of the entropy of $p_Y(\cdot)$ results

$$\hat{H}_1(Y) = \frac{1}{N}\sum_{i=1}^{N}\log\left\{N\left[TY_i - Y_i\right]\right\}. \tag{3.3}$$

This estimate is the average of the log of the distance between adjacent random variables in the amplitude-ordered time series. This estimate is similar to that found in [21].

As there is no equivalent ordering of elements in $\Re^d$, for $d > 1$, a more general setting for the estimate in equation (3.3) must be obtained. The estimate in (3.3) can also be viewed in the following manner: the random variables in the time series, $\{Y_n\}$ are placed on the real line, $\Re^1$. Consider an arbitrary random variable Y_i; searching for the nearest larger random variable, TY_i, on $\Re^1$ can be visualized as an unidirectional expanding sphere which stops expanding as soon as the point TY_i is encountered. The first order estimate is then the average of the logarithm of the volume of the largest sphere centered at Y_i that contains no random variable larger than itself. This view of the first order entropy estimate provides some insight as to how higher order joint entropy estimates can be obtained. The sample space points $Y_i \in \Re^d$ are created from the original time series $\{Y_n\}$ in the following manner

$$Y_1 = \begin{bmatrix} Y_1 \\ \vdots \\ Y_d \end{bmatrix}, \qquad Y_2 = \begin{bmatrix} Y_{d+1} \\ \vdots \\ Y_{2d} \end{bmatrix}, \cdots\cdots, Y_{N_d} = \begin{bmatrix} Y_{d(N_d-1)+1} \\ \vdots \\ Y_{dN_d} \end{bmatrix} \tag{3.4}$$

where $N_d = \left\lfloor \frac{N}{d} \right\rfloor$. The following relationship holds for the random vector $Y \in D \subset \Re^d$

$$\int_D p_Y(y)\, dV = P_Y(y)$$

where D is the volume over which integration is performed. Construct an expanding sphere (a generalized sphere in $\Re^d$) centered at an arbitrary data point, Y_i; expand the sphere until it encounters another data point (this will be termed as the *nearest neighbor* of Y_i) denoted as $Y_{i,NN}$. Letting this open sphere correspond to the region D, the volume it encloses is $V_{i,NN}$. When $V_{i,NN}$ is small, the above integral can be approximated by

$$P_Y(y) \approx p_Y(Y_i)\, V_{i,NN}.$$

Using the empirical distribution function one obtains

$$p_Y(Y_i)\, V_{i,NN} \approx \frac{1}{N_d}, \tag{3.5}$$

where N_d is the total number of sample space points, $Y_i \in \Re^d$, created from the observed time series $\{Y_n\}$. As in the first order entropy estimate the average value of $-\log p_Y(\cdot)$ would serve as an estimate of the $E\left[-\log p_Y(\cdot)\right]$, i.e., the joint entropy (see equation (2.4)) of the random variables $\{Y_i, Y_{i-1}, \ldots, Y_{i-d+1}\}$. Using equation (3.5) one obtains the following d^{th} order entropy estimate

$$\hat{H}_d(Y) = \frac{1}{N_d} \sum_{i=1}^{N_d} \log \left[N_d\, V_{i,NN}\right]. \tag{3.6}$$

As $N_d = \left\lfloor \frac{N}{d} \right\rfloor$, N_d decreases with d; thus given a time series $H_d(Y)$ will have to be estimated from fewer elements for larger values of d.

As discussed in § 2.3 (see discussion following equation (2.10)) the model order of the time series, p, is given by the smallest value of l for which the quantity $H(Y_i \mid Y_{i-1}, \ldots, Y_{i-l}) - H(Y_i \mid Y_{i-1}, \ldots, Y_{i-l-1})$ is zero. If the l^{th}-order conditional entropy is estimated by $\hat{H}(Y_i \mid Y_{i-1}, \ldots, Y_{i-l})$ $(= \hat{H}_{l+1}(Y) - \hat{H}_l(Y))$ then the model order is given by the value of m for which the difference $\hat{H}(Y_i \mid Y_{i-1}, \ldots, Y_{i-m}) - \hat{H}(Y_i \mid Y_{i-1}, \ldots, Y_{i-m-1})$ (denoted by ΔH_m) satisfies the null hypothesis for all $m \geq l$ and does not satisfy the null hypothesis for any $m < l$. To test the hypotheses $\Delta H_m = 0$ and $\Delta H_m \neq 0$ the distribution of this statistic under null hypothesis is required. As it was difficult to derive the distribution of ΔH_m analytically, a normal distribution was assumed. With this assumption one requires only the bias and variance of ΔH_m to test it for null hypothesis. With this goal in mind we now obtain the bias and variance of the entropy estimate $\hat{H}_d(Y)$.

3.2. Bias of Entropy Estimate

To obtain the bias of the first order entropy estimate, $\hat{H}_1(Y)$ one needs to obtain the bias of $E\log\left[TY_n - Y_n\right]$. Consider

$$E\log\left[TY_n - Y_n\right] = E\left\{E\left[\log\,(TY_n - Y_n)\,|\,Y_n\right]\right\}.$$

To simplify the right hand side of the above equation one needs the conditional distribution of TY_n given Y_n; the following theorem from [13] which provides a link between the order statistics of a set of independent and identically distributed random variables and a nonhomogeneous Poisson process.

> Let $\{N(t), t \geq 0\}$ be a Poisson process with intensity function $\lambda(t)$. Under the condition that $N(T) = k$, the k occurence times $\tau_1 < \tau_2 < \cdots < \tau_k$ in the interval $(0, T]$ at which events occur are random variables having the same joint distribution as if they were the order statistics corresponding to k independent random variables $U_1, \ldots, U_k$ having the common distribution function
>
> $$P_{U_j}(u) = \frac{\int_0^u \lambda(t)\,dt}{\int_0^T \lambda(t)\,dt}, \quad 0 \leq u \leq T. \tag{3.7}$$

Under the assumption that the time series $\{Y_n\}$ is comprised of independent and identically distributed random variables the probability density function of the difference, $TY_n - Y_n = \gamma$ conditioned on Y_n is (by the above theorem) [19]

$$p_{\gamma|Y_n}(x\,|\,y) = \lambda(x+y)\,e^{-\int_y^{y+x} \lambda(\alpha)\,d\alpha}.$$

Using the above conditional density function in the identity we have

$$E\log\left[TY_n - Y_n\right] = \int_{-\infty}^{\infty} p_{Y_n}(x) \int_0^{\infty} \log\gamma\,\lambda(x+\gamma)\, e^{-\int_x^{x+\gamma}\lambda(\alpha)\,d\alpha}\, d\gamma\, dx.$$

Denote the integral $\int_0^x \lambda(\alpha)\,d\alpha$ by $\Lambda(x)$. Making the transformation $y = \Lambda(x+\gamma) - \Lambda(x)$ we obtain

$$E\log\left[TY_n - Y_n\right] = \int_{-\infty}^{\infty} p_{Y_n}(x) \int_0^{\infty} \log\left[\Lambda^{-1}\left[y + \Lambda(x)\right] - x\right] e^{-y}\, dy.$$

Using a first order Taylor's approximation about $y = 0$ for $\Lambda^{-1}(y+\Lambda(x)) - x$ one obtains

$$E\log\left[TY_n - Y_n\right] = \int_{-\infty}^{\infty} p_{Y_n}(x) \int_0^{\infty}\log\left[\frac{y}{\lambda(x)}\right] e^{-y}\, dy.$$

Simplifying the above expression using results from [5] one obtains

$$E\log\left[TY_n - Y_n\right] = -\int_{-\infty}^{\infty} p_{Y_n}(x)\log\lambda(x)\, dx - C. \tag{3.8}$$

where C is Euler's constant.

If there are N points in the interval $(0, T]$ i.e., $N(T) = N$, then clearly $\Lambda(T) = EN(T) = N$ (as in the theorem we are concerned with a Poisson process conditioned on $N(T) = N$). Using this result and equation (3.7) we have the following relation

$$\lambda(x) = N\, p_{Y_n}(x).$$

Using the above relation in equation (3.8) one gets

$$E\log\left[TY_n - Y_n\right] = -\int_{-\infty}^{\infty} p_{Y_n}(x)\log p_{Y_n}(x)\, dx - \log N - C.$$

Using the above result in equation (3.3) one obtains $E\hat{H}_1(Y) = H_1(Y) - C$. Thus $\tilde{H}_1(Y) = \hat{H}_1(Y) + C$ is asymptotically an unbiased estimator of $H_1(Y)$.

Simulations were performed to test the result that $\tilde{H}_1(Y)$ was a distribution-free statistic. A time series with independent and identically distributed random variables was formed. Simulations were performed for the Exponential, Laplacian, Normal, Rayleigh and Uniform densities; the probability density functions and the mathematical expressions for the corresponding first order entropies are included in Table 3.1.

Table 3.1

	$p(y)$	$H_1(Y)$
Exponential	$\lambda e^{-\lambda y},\ y>0$	$1-\log\lambda$
Laplacian	$\frac{\lambda}{2}e^{-\lambda\lvert y\rvert},\ -\infty<y<\infty$	$1-\log\left[\frac{\lambda}{2}\right]$
Normal	$\frac{1}{\sqrt{2\pi}\,\lambda}e^{\frac{-y^2}{2\lambda^2}},\ -\infty<y<\infty$	$\frac{1}{2}\log(2\pi e\lambda^2)$
Rayleigh	$\frac{y}{\lambda^2}e^{-\frac{y^2}{2\lambda^2}},\ y>0$	$1+\frac{1}{2}\left[C-\log\left[\frac{2}{\lambda^2}\right]\right]$
Uniform	$\frac{1}{\lambda},\ -\frac{\lambda}{2}<y<\frac{\lambda}{2}$	$\log\lambda$

(C is Euler's constant)

The results of the simulations are tabulated in Table 3.2. The table includes the theoretical value of the entropy $H_1(Y)$. The mean and the standard deviation of the entropy estimate $\tilde{H}_1(Y)$ based on 100 repetitions are also included in the table. The estimate deviates no more than 6 percent from the theoretical value for any distribution. From the table it is also clear that the performance of the estimator $\tilde{H}_1(Y)$ does not depend strongly on the probability density function.

Table 3.2

$p_Y(y)$	λ	$H_1(Y)$	$\bar{H}_1(Y)$	
			Mean	SD
Exponential	1	1.0000	0.9951	0.0285
Laplacian	$\sqrt{2}$	1.3466	1.3441	0.0312
Normal	1	1.4189	1.4159	0.0241
Rayleigh	$\sqrt{2/\pi}$	0.7162	0.7180	0.0253
Uniform	$\sqrt{12}$	1.2425	1.2431	0.0169

($p_Y(\cdot)$ denotes the probability density function of $Y_n \in \{Y_n\}$; λ is its parameter. The Mean and SD (standard deviation) were computed on basis of 100 runs. $N = 2000$)

A similar analysis of the estimate of higher order ($d > 1$) entropy, $\hat{H}_d(Y)$, (see equation (3.6)) was difficult; hence the bias was estimated in an empirical manner. The bias was assumed to be equal to that derived in one dimension. The estimate $\bar{H}_d(Y) = \hat{H}_d(Y) + C$ is a natural extension of the unbiased (asymptotically) estimator of the first order joint entropy, $\bar{H}_1(Y)$. The performance of this estimate was studied through simulations for the sample cases $d = 2, 3$. l_2 distance measure was used in computing the volume, $V_{i,NN}$, in equation (3.6). The simulations were performed on a time series with independent and identically distributed random variables; the Exponential, Laplacian, Normal, Rayleigh and Uniform densities were considered. From equation (2.9) one has, for a time series $\{Y_n\}$ with independent and identically distributed (i.i.d.) random variables, $H_d(Y) = d\,H_1(Y_n)$. $Y \in \Re^d$ were constructed as described in § 3.1 (see equation (3.4)) and hence the theoretical value of the joint entropy for $d > 1$ is known. The results of the simulations are tabulated in Table 3.3. The mean and standard deviation of the entropy estimate $\bar{H}_d(Y)$ (based on 100 repetitions) indicate that the estimate deviates no greater than 6 percent from the theoretical value.

Table 3.3

$p_Y(y)$	λ	$H_2(Y)$	$\bar{H}_2(Y)$		$H_3(Y)$	$\bar{H}_3(Y)$	
			Mean	SD		Mean	SD
Exponential	1	2.0000	2.0124	0.0478	3.0000	3.0617	0.0630
Laplacian	$\sqrt{2}$	2.6932	2.6897	0.0148	4.0398	4.0220	0.0537
Normal	1	2.8378	2.8348	0.0389	4.2567	4.2382	0.0433
Rayleigh	$\sqrt{2/\pi}$	1.4324	1.4375	0.0366	2.1486	2.1431	0.0418
Uniform	$\sqrt{12}$	2.4850	2.4983	0.0323	3.7275	3.7952	0.0320

($p_Y(\cdot)$ denotes the probability density function of $Y_n \in \{Y_n\}$; λ is its parameter.
The Mean and SD (standard deviation) are based on 100 runs. $N_d = 2000$ for $d=2, 3$)

The results indicate that the conjectured bias was appropriate. Further, the simulations indicate that the estimate does not depend strongly on the underlying distribution.

3.3. Variance of Entropy Estimate

An analytical expression will now be obtained for the variance of the first order entropy estimate, $\hat{H}_1(Y)$ for a time series of independent random variables. Using the notation and approach of determining the first moment of log $\left[TY_n - Y_n\right]$ we have the following expression for the second moment

$$E\left\{\log(TY_n - Y_n)\right\}^2 = \int_{-\infty}^{\infty} p_{Y_n}(x) \int_0^{\infty} (\log\gamma)^2 \, \lambda(x+\gamma) \, e^{-\int_x^{x+\gamma} \lambda(\alpha)\, d\alpha} \, d\gamma \, dx.$$

A first order Taylor approximation for $\Lambda^{-1}(y+\Lambda(x)) - x$ about $y = 0$ results in

$$E\left[\log(TY_n - Y_n)\right]^2 = \int_{-\infty}^{\infty} p_{Y_n}(x) \int_0^{\infty} \left[\log\left[\frac{y}{\lambda(x)}\right]\right]^2 e^{-y}\, dy.$$

Simplifying the above expression using results from [5] we obtain

$$= \frac{\pi^2}{6} + C^2 + 2C \int_{-\infty}^{\infty} p_{Y_n}(x) \log\lambda(x)\, dx + \int_{-\infty}^{\infty} p_{Y_n}(x) \left[\log\lambda(x)\right]^2 dx$$

where C is the Euler's constant. Using the above equation in conjunction with equation (3.8) and simplifying one obtains the following expression for the variance of log $\left[TY_n - Y_n\right]$

$$Var\left\{\log \left[TY_n - Y_n\right]\right\} = \frac{\pi^2}{6} + Var\left[\log p_{Y_n}(x)\right].$$

We will assume that log $\left[TY_i - Y_i\right]$ and log $\left[TY_j - Y_j\right]$ are independent and identically distributed for all $i \neq j$; this assumption is exact only when Y_n is uniformly distributed. Using the above result one has the following approximate variance of the first order entropy estimate $\hat{H}_1(Y)$

$$Var\hat{H}_1(Y) = \frac{\pi^2}{6N} + \frac{1}{N}\, Var\left[\log p_Y(x)\right]. \qquad (3.9)$$

The expressions for the above variance expression for different distributions are given in Table (3.4). ,

Table 3.4	
	$p_Y(y)$
Exponential	$\frac{\pi^2}{6N}+\frac{1}{N}$
Laplacian	$\frac{\pi^2}{6N}+\frac{1}{N}$
Normal	$\frac{\pi^2}{6N}+\frac{1}{2N}$
Rayleigh	$\frac{\pi^2}{6N}+\frac{\pi^2}{24N}$
Uniform	$\frac{\pi^2}{6N}$

The following the statistic was used to estimate the variance of $\tilde{H}_1(Y)$

$$\overline{Var}\left[\tilde{H}_1(Y)\right]=\frac{1}{N}\sum_{i=1}^{N}\left\{\log N\left[TY_i-Y_i\right]\right\}^2-\left\{\frac{1}{N}\sum_{i=1}^{N}\log N\left[TY_i-Y_i\right]\right\}^2. \tag{3.10}$$

Simulations were performed to demonstrate that the above statistic estimated the variance expression obtained in equation (3.9). Simulations were performed on time series with independent and identically distributed random variables for the Exponential, Laplacian, Normal, Rayleigh and Uniform densities; these densities are given in Table (3.1). The results of these simulations are tabulated in Table (3.5).

Table 3.5

$p_Y(y)$	λ	$Var\,\tilde{H}_1(Y)$	$\overline{Var}\,\tilde{H}_1(Y)$	
			Mean	SD
Exponential	1	1.323×10^{-3}	1.307×10^{-3}	4.4×10^{-5}
Laplacian	$\sqrt{2}$	1.323×10^{-3}	1.310×10^{-3}	4.7×10^{-5}
Normal	1	1.072×10^{-3}	1.061×10^{-3}	4.3×10^{-5}
Rayleigh	$\sqrt{2/\pi}$	1.028×10^{-3}	1.020×10^{-3}	4.3×10^{-5}
Uniform	$\sqrt{12}$	0.823×10^{-3}	0.829×10^{-3}	3.8×10^{-5}

($p_Y(\cdot)$ denotes the probability density function of $Y_n \in \{Y_n\}$; λ is its parameter. The Mean and SD (standard deviation) were computed on basis of 100 runs. $N = 2000$)

The simulations indicate that the variance estimate of $\tilde{H}_1(Y)$ deviates no more than 8 percent from the theoretical value. Further, it is also clear that the performance of the variance estimate does not depend strongly on the density function.

A similar analysis of the variance of higher order ($d > 1$) entropy estimates, $\tilde{H}_d(Y)$ (equation (3.6)) was difficult; hence the following empirical variance estimate - a straight forward extension of equation (3.10) - was used.

$$\overline{Var}\tilde{H}_d(Y) = \frac{1}{N_d} \sum_{i=1}^{N_d} \left[\log N_d\, V_{i,NN}\right]^2 - \left[\frac{1}{N_d} \sum_{i=1}^{N_d} \log N_d\, V_{i,NN}\right]^2 . \tag{3.11}$$

where L_2 distance measure was used in computing the volume $V_{i,NN}$. Simulations were performed on a time series with independent and identically distributed random variables; the Exponential, Laplacian, Normal, Rayleigh and Uniform densities were considered. The simulations were performed for the sample cases $d = 2, 3$. The variance was estimated with the expression in equation (3.11). The results of the simulations are tabulated in Table (3.6) The results of the simulations were compared with the following extension of the theoretical result of equation (3.9)

$$Var\tilde{H}_d(Y) = \frac{\pi^2}{6N_d} + \frac{1}{N_d}\, Var\left[\log p_Y(y)\right]. \tag{3.12}$$

When $Y_i \in \Re^d$ created from the time series (as described in 4, equation (3.4)) are independent and identically distributed the above equation reduces to

$$Var\tilde{H}_d(Y) = \frac{\pi^2}{6N_d} + \frac{d}{N_d}\, Var\left[\log p_Y(y)\right].$$

The numerical values of the above expression for the corresponding simulations are also included in Table (3.6) and a comparison with the variance estimate indicates that they are not significantly different. The results indicate that the theoretical variance results obtained for the first order entropy estimate, $\tilde{H}_1(Y)$, although difficult to derive analytically, extend to higher order entropy estimates.

Table 3.6

$p_Y(y)$	λ	$Var\,\tilde{H}_2(Y)$ $\times 10^{-3}$	$\overline{Var}\,\tilde{H}_2(Y)$		$Var\,\tilde{H}_3(Y)$ $\times 10^{-3}$	$\overline{Var}\,\tilde{H}_3(Y)$	
			Mean $\times 10^{-3}$	SD $\times 10^{-5}$		Mean $\times 10^{-3}$	SD $\times 10^{-5}$
Exponential	1	1.822	1.822	8.7	2.323	2.317	9.4
Laplacian	$\sqrt{2}$	1.822	1.803	8.3	2.323	2.224	8.2
Normal	1	1.322	1.295	6.0	1.573	1.476	7.5
Rayleigh	$\sqrt{2/\pi}$	1.234	1.233	5.9	1.439	1.419	7.7
Uniform	$\sqrt{12}$	0.823	0.838	4.7	0.823	0.860	6.3

($p_Y(\cdot)$ denotes the probability density function of $Y_n \in \{Y_n\}$; λ is its parameter.
The Mean and SD (standard deviation) are based on 100 runs. $N_d = 2000$ for d=2, 3)

3.4. Estimating Model Order

The model order of any given time series will be estimated by the procedure outlined in § 3.1; the quantity ΔH_l will be computed for increasing values values of l. ΔH_l will then be tested for the null hypothesis. As ΔH_l is the difference between the l^{th}-order and the $(l+1)^{th}$-order conditional entropies and the l^{th}-order conditional entropy can be written as the difference between the $(l+1)^{th}$-order and l^{th}-order joint entropies one obtains

$$\Delta H_l = -\hat{H}_{l+2}(Y) + 2\,\hat{H}_{l+1}(Y) - \hat{H}_l(Y). \tag{3.13}$$

where $\hat{H}_l(Y)$ is as defined in equation (3.3) for $l = 1$ and in equation (3.6) for $l \geq 2$. Using results of § 3.2 one easily sees that ΔH_l is an unbiased (asymptotically) estimator of the difference in the l^{th}-order and $(l+1)^{th}$-order conditional entropies. With the assumption that $\hat{H}_{d_1}(Y)$ and $\hat{H}_{d_2}$ are statistically independent (this was verified through simulations) for any $d_1 = d_2$ we obtain (using equation (3.13))

$$V\hat{a}r\Delta H_l = \overline{Var}\hat{H}_{l+2}(Y) + 4\,\overline{Var}\hat{H}_{l+1}(Y) + \overline{Var}\hat{H}_l(Y) \tag{3.14}$$

where $\overline{Var}(\cdot)$ is as defined in equation (3.10) for $d = 1$ and in equation (3.11) for $d \geq 2$. With the foregoing assumptions ΔH_l satisfies the null hypothesis if it lies in the range $[-2\,\beta, 2\,\beta]$ where $\beta = \sqrt{V\hat{a}r\Delta H_l}$.

To assess the characteristics of the new model order determination procedure a series of simulations were performed on data with a first order dependence structure. The performance measure used was the model order selected in repeated simulations (runs). The results of simulations with a first order Gaussian AR model are tabulated in Table 3.7; for a first order coefficient larger in magnitude than 0.5 the entropy method selects the correct model order in greater than 90 percent of the runs. The entropy method is clearly not very sensitive to low levels of dependence in a time series whose elements are jointly normal.

Table 3.7
Performance of Entropy Method on First Order AR Time Series

	Percentage of Correct Order Prediction	
ρ	Gaussian	Exponential
-0.1	0 (0.02, 0.20)	29 (0.31, 0.48)
-0.2	2 (0.02, 0.14)	95 (0.99, 0.22)
-0.3	7 (0.07,0.26)	99 (1.01, 0.10)
-0.4	37 (0.41, 0.53)	96 (1.04, 0.20)
-0.5	90 (0.98, 0.32)	94 (1.06, 0.24)
-0.6	93 (1.07, 0.26)	94 (1.06, 0.24)
-0.7	97 (1.03, 0.17)	96 (1.04, 0.20)
-0.8	96 (1.04, 0.20)	98 (1.02, 0.14)
-0.9	95 (1.05, 0.22)	96 (1.04, 0.20)

(ρ ($\equiv \theta_1$) is the coefficient of the first order AR time series. N = 6000. Percentages based on 100 simulations. The numbers in brackets are the mean and standard deviation of the model order chosen in the 100 runs)

The results of the performance of the three techniques (MAICE, MDL and Entropy Method) on Gaussian AR time series are plotted in figure 3.1b; the percentage of correct model order prediction is plotted against the magnitude of first order coefficient (the values are taken from Table 2.1 and Table 3.7). Clearly the performance of the Entropy method on a first order Guassian AR time series is superior to the performance of MAICE technique for first order coefficients larger in magnitude than 0.5; the trend is reversed for values of coefficients smaller in magnitude than 0.5. The MDL technique selects the model order far more accurately than the Entropy method for small values of coefficients (less in magnitude than 0.5) and the two methods have comparable performance levels for larger values of coefficients.

The Entropy method performs very well on an Exponential AR time series; for a first order coefficient larger in magnitude than 0.1 the method selects the model order correctly in greater than 94 percent of the runs (the results are tabulated in Table 3.7). The method is far more sensitive to small values of coefficients in an Exponential AR time series than in a Gaussian AR time series.

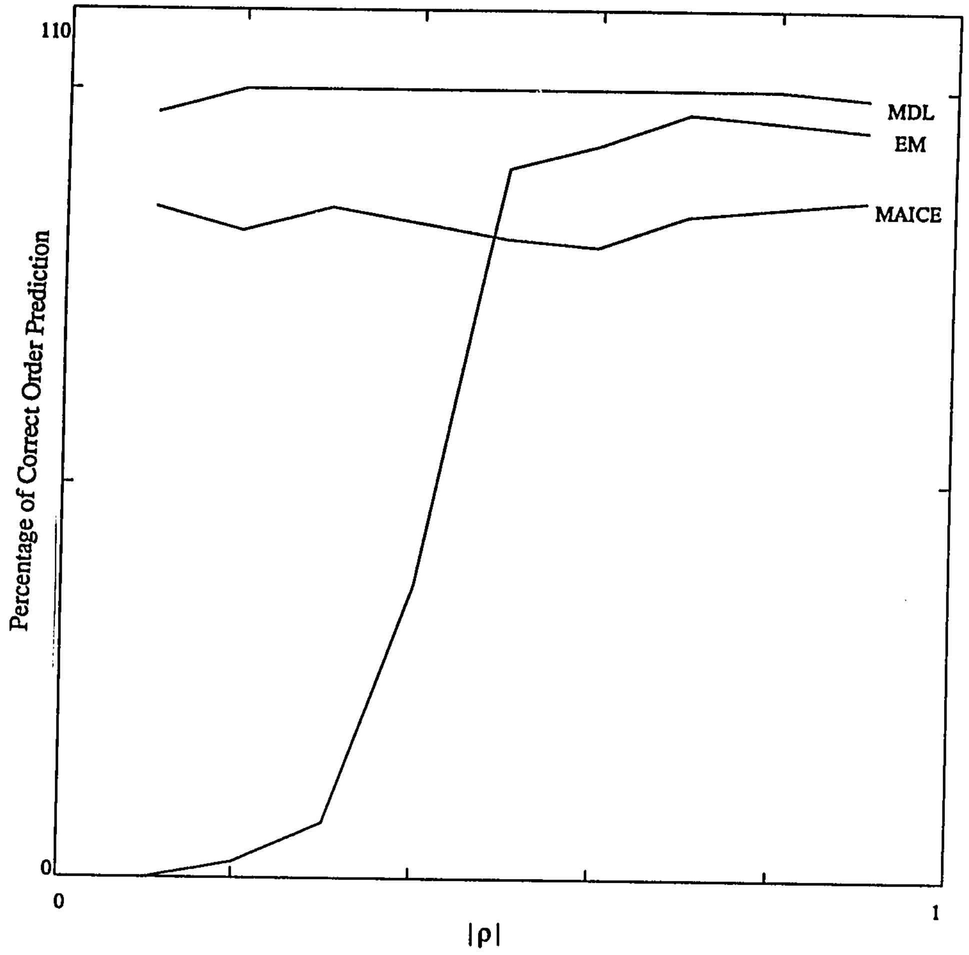


Figure 3.1a: Performance of MAICE, MDL and Entropy Method (EM) on Gaussian AR time series. The percentage of correct model order selected (based on 100 runs) is plotted against the *magnitude* of the first order coefficient, ρ; the percentages and the corresponding first order coefficients are as in Table 2.1 (for MAICE and MDL) and in Table 3.7 (for Entropy Method).

The results of the performance of the three techniques (MAICE, MDL and Entropy Method) on Exponential AR time series are plotted in figure 3.1b; the percentage of correct model order prediction is plotted against the magnitude of first order coefficient (the values are taken from Table 2.1 and Table 3.7). It is evident that the Entropy Method and the MDL techniques have similar performance levels on an Exponential AR time series; these two methods select the model order more accurately than the MAICE technique.

The performance of the Entropy Method on data with a nonlinear dependence structure was studied through simulations. A first order time series with an exponential (nonlinear) dependence structure (see equation (2.1)) was generated by the technique discussed by Johnson and

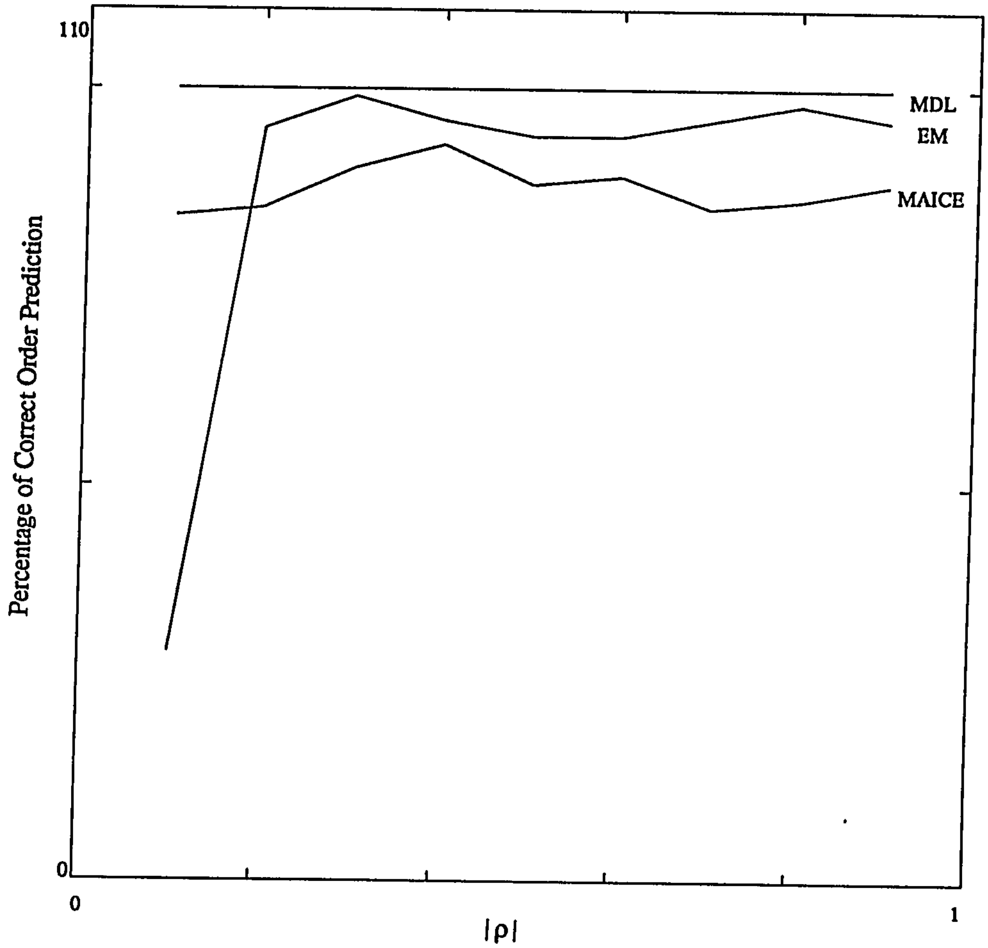


Figure 3.1b: Performance of MAICE, MDL and Entropy Method (EM) on Exponential AR time series. The percentage of correct model order selected (based on 100 runs) is plotted against the *magnitude* of the first order coefficient, ρ; the percentages and the corresponding first order coefficients are as in Table 2.1 (for MAICE and MDL) and in Table 3.7 (for Entropy Method).

Linebarger [7]. The method has been outlined in § 2.1. The results of the simulations have been tabulated in Table 3.8. The serial correlation coefficient, ρ (which is the first order coefficient under the assumption of an AR model), was computed for the data and has been included in the table. The results indicate that the Entropy Method selects the correct model order higher than 96 percent for moderately large to large dependence. The Entropy Method is not very sensitive to dependence levels corresponding to $\theta_{1,1} \geq 0.002$

Table 3.8
Performance of Entropy Method on a Time Series with First Order Nonlinear Dependence

$\theta_{1,1}$	ρ	Percentage of Correct Order Prediction
.002	-0.12	35 (0.41, 0.55)
.003	-0.17	81 (0.87, 0.42)
.006	-0.30	98 (1.02, 0.14)
.009	-0.40	97 (1.03, 0.17)
.012	-0.51	96 (1.04, 0.20)
.015	-0.59	96 (1.04, 0.20)
.02	-0.69	96 (1.04, 0.20)

(ρ is the serial correlation coefficient. $\theta_{1,2}$ was held constant at 195.4. N = 6000. Percentages based on 100 simulations. The numbers in brackets are the mean and standard deviation of the model order chosen in the 100 runs)

The results of the performance of the three techniques (MAICE, MDL and Entropy Method) on Exponential AR time series are plotted in figure 3.1c; the percentage of correct model order prediction is plotted against the magnitude of first order coefficient (the values are taken from Table 2.2 and Table 3.8). It is evident that the performance of the Entropy Method improves with increasing nonlinear dependence, whereas the performance of MAICE and MDL depreciates with increasing dependence (nonlinear); MAICE and MDL do poorly even on data with moderately large nonlinear dependence.

Evidently the Entropy Method performs poorly on a Gaussian AR model for small first order coefficients. This poor performance is due to the low sensitivity of the conditional entropy of Gaussian AR time series to first order coefficients; applying the expression for joint entropy

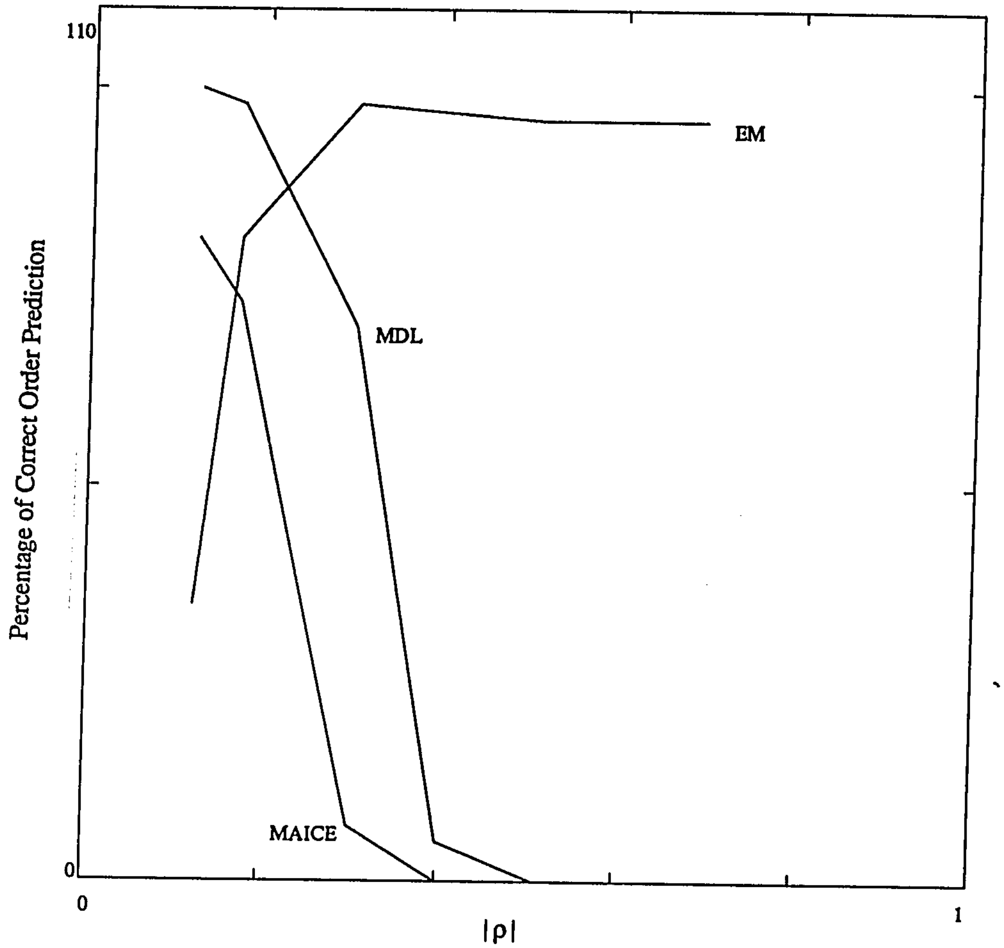


Figure 3.1c: Performance of MAICE, MDL and Entropy Method (EM) on time series with nonlinear dependence structure. The percentage of correct model order selected (based on 100 runs) is plotted against the *magnitude* of the serial correlation coefficient, ρ; the percentages and the corresponding first order coefficients are as in Table 2.2 (for MAICE and MDL) and in Table 3.8 (for Entropy Method).

(equation (2.5)) in equation (2.7) one obtains the expression for conditional entropy of a random variable X given the random variable Y, $H(X \mid Y)$, and hence the expression for $H(X \mid Y) - H(X)$ as $\frac{1}{2}\log(1-\rho^2)$ (ρ is the serial correlation coefficient). $H(X \mid Y) - H(X)$ is plotted against ρ in figure 3.2; the low sensitivity of this difference to the coefficient is evident from the figure. The

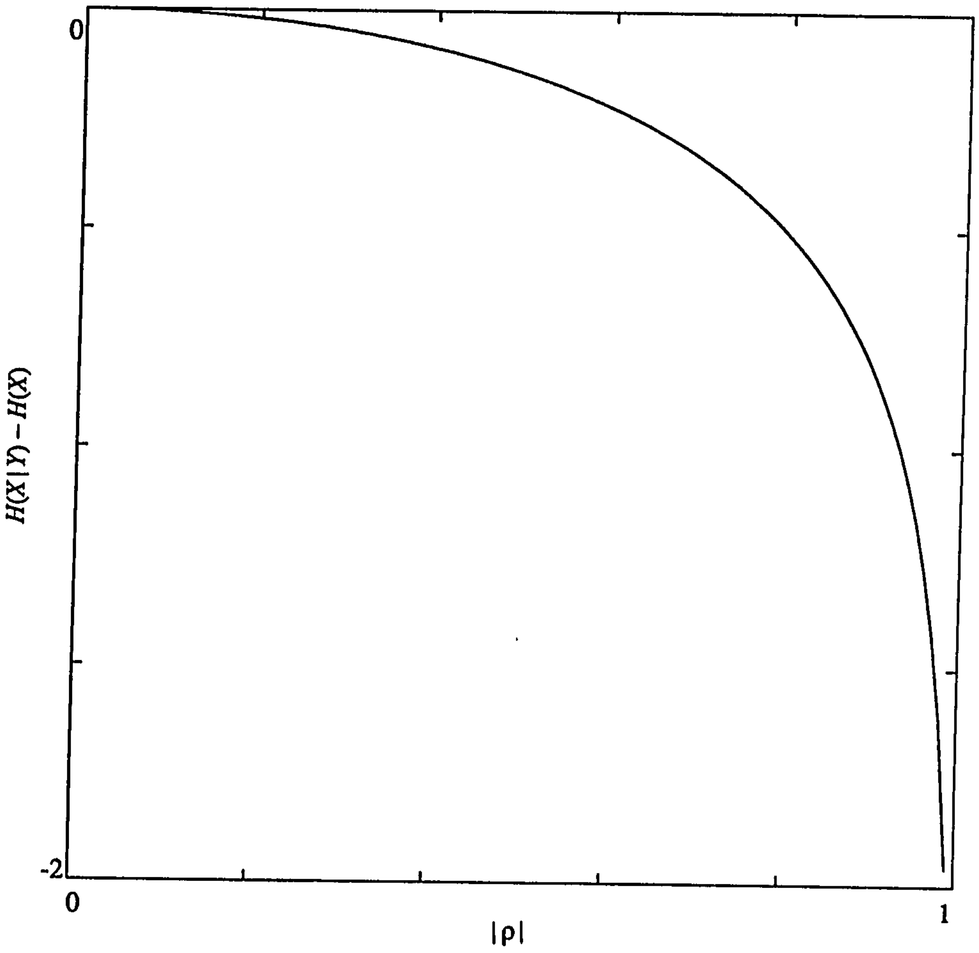


Figure 3.2: H(X|Y) - H(X) is plotted against the serial correlation coefficient, ρ for a first order Gaussian AR time series; $H(X \mid Y) - H(X) = \frac{1}{2}\log(1 - \rho^2)$.

Entropy Method performs well on first order Exponential AR time series; the method is sensitive to low levels of dependence. The performance of the method on time series with nonlinearly dependence structure is good for moderately large to large dependence; the method is not sensitive to low dependence levels.

3.5. Size of Data required for analysis

Simulations were performed on first order Exponential AR time series to determine the minimum length of the time series (value of N) that is required to obtain a performance level comparable to that for $N = 6000$ (see Table 3.8). The simulations were performed for $N = 100$, 1000, 2000 and 3000. The results are tabulated in Table 3.9.

Table 3.9					
	Percentage of Correct Order Prediction				
ρ	$N = 100$	$N = 1000$	$N = 2000$	$N = 3000$	$N = 6000$
-0.1	0	1	2	10	29
-0.2	1	15	42	73	95
-0.3	0	41	83	99	99
-0.4	2	79	100	98	96
-0.5	4	97	96	99	94
-0.6	16	99	100	98	94
-0.7	28	97	97	97	96
-0.8	55	94	97	93	98
-0.9	78	93	94	92	96

(ρ is the coefficient of the first order Exponential AR time series. Percentages based on 100 simulations.)

From Table 3.9 it is evident that one needs over 3000 elements in the time series to estimate the model order with a performance level comparable to that for $N = 6000$; this is the amount of data required for a time series with *first order* dependence. The situation is likely to get worse for time series with higher model orders. This observation is not surprising for three reasons: firstly estimating the model order of a p^{th}-order time series involves estimating the $(p+2)^{th}$-order joint entropy (see § 2.3). The second reason is that an N-element time series results in only $\left\lfloor \frac{N}{d} \right\rfloor$ elements in $\Re^d$. Thus one would have to estimate the $(p+2)^{th}$ joint entropy from only $\frac{N}{p+2}$ elements. The third reason could be termed as the "curse of dimensionality", whereby the number of elements required to fill a unit volume in space increases exponentially with dimension. Thus the Entropy Method is a data-intensive model order determining

3.6. Computational Complexity

The unbiased (asymptotically) first order entropy $\tilde{H}_1(Y)$ is obtained by first computing the estimate $H_1(Y)$ and then adjusting for the bias. The quantity $H_1(Y)$ (see equation (3.3)) is computed by sorting the given time series by magnitude and then averaging the logarithm of the difference between adjacent elements (of this amplitude-ordered time series). Most of the computational complexity is associated with amplitude-ordering the time series; if an efficient sorting algorithm, for example quicksort, is used the complexity is of the order $O(N\log N)$.

The key to computing the d^{th}-order entropy, $H_d(Y)$ $(d > 1)$ is computing the L_2 distance to the nearest neighbor of each of the elements in $\Re^d$ (see equation (3.6)). If a brute force approach (of computing the distances to all other points and then taking the minimum of these distances) is used to locate the nearest neighbor the order of computational complexity (in terms of number of distances that have to be computed) is $O(d\,N_d^2)$. Thus for $N_d = 2000$ the complexity is of the order 10^6. Thus the time required for computing $H_d(Y)$ is large and hence the computational resources required is enormous. The computational requirement is further increased by the fact that, as discussed in § 3.5, the Entropy Method is data-intensive.

3.7. Estimating entropy of quantized data

The recording of data is constrained by the fact that computing machines have finite precision; hence the recorded variable of a time series is the quantized value (see figure 3.3a) of the true variable (which may be, for example, amplitude or time). The quantization process can be modeled as passing the time series $\{Y_n\}$, through a finite time integrator, then sampling its output (see figure 3.3b); this model is appropriate for determining the probability density function of the quantized data. The impulse response of this integrator, $h_\Delta(y)$, is given by

$$h_\Delta(y) = \begin{bmatrix} 1 & 0 \leq y \leq \Delta \\ 0 & \text{elsewhere} \end{bmatrix}$$

The output of the finite time integrator is sampled every Δ seconds resulting in the discrete random variable Y' that takes on the values Y_n with probability p_n; p_n is given by

$$p_n = \left[p_Y(y) \otimes h_\Delta(y) \right|_{y=n\Delta}$$

where $p_Y(\cdot)$ is probability density function of the input to the integrator and $\otimes$ denotes convolution. Simplifying one obtains

$$= \left[\int_{-\infty}^{\infty} p_Y(\alpha)\, h_\Delta(y-\alpha)\, d\alpha \right|_{y=n\Delta}$$

$$= \left[\int_{y-\Delta}^{y} p_Y(\alpha)\, d\alpha \right|_{y=n\Delta}$$

$$= \int_{(n-1)\Delta}^{n\Delta} p_Y(\alpha)\, d\alpha.$$

Applying the Mean Value Theorem one obtains

$$p_n = \Delta\, \hat{p}_Y(n\Delta) \tag{3.15}$$

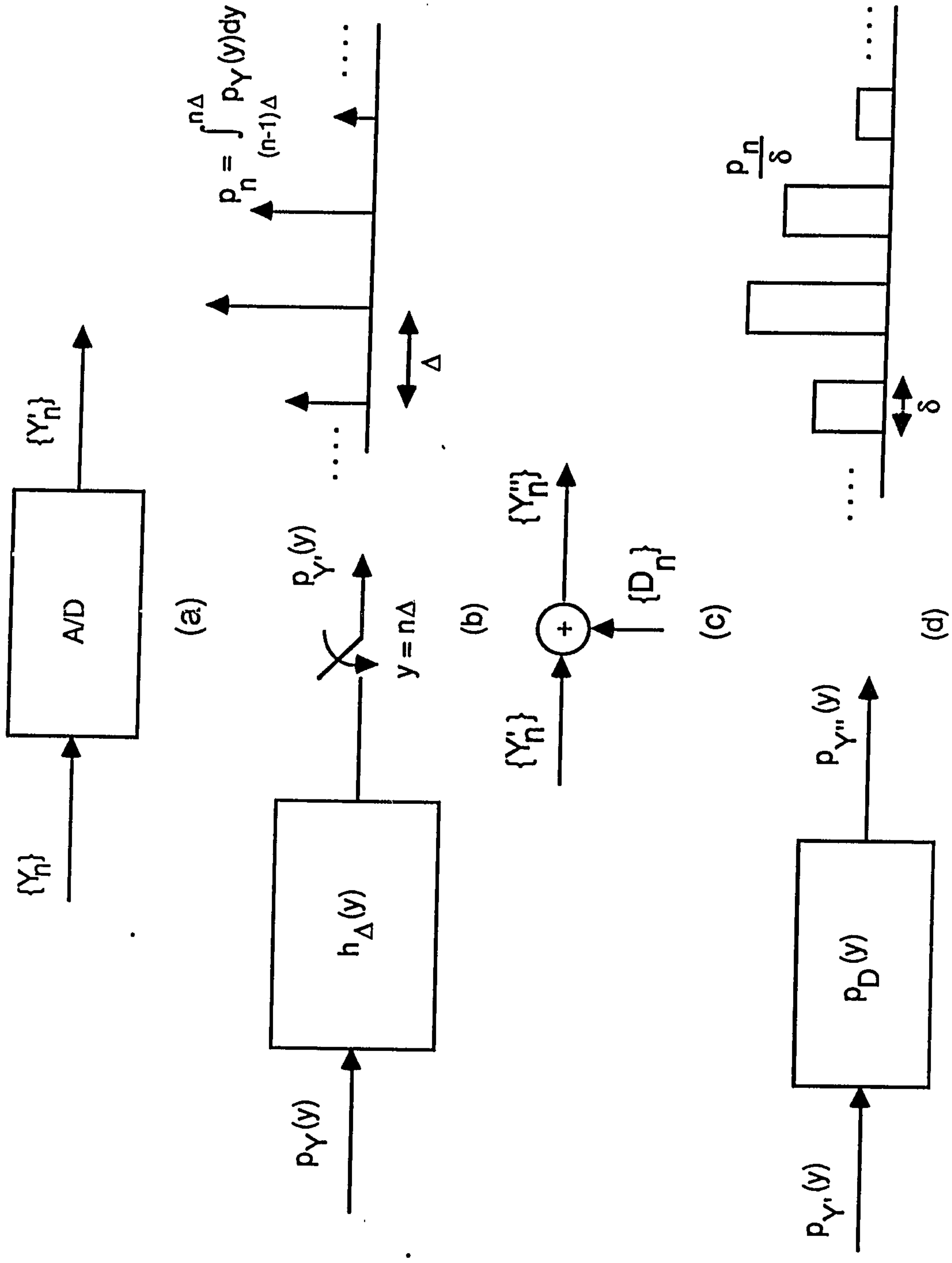
Figure 3.3: **Panel a** describes the quantization process of passing the time series $\{Y_n\}$ through an Analog to Digital (A/D) converter; $\{Y'_n\}$ is the output of the A/D converter and Y'_n is the quantized value of Y_n. **Panel b** is a block diagram modeling the quantization procedure. The probability density function of the time series $\{Y_n\}$, $p_Y(y)$, is passed through a finite time integrator whose impulse response is given by

$$h_\Delta(y) = \begin{cases} 1 & 0 \le y \le \Delta \\ 0 & \text{elsewhere} \end{cases}.$$

The output of the integrator is sampled every Δ seconds and the resulting density function $p_{Y'}(y)$ is an impulse train; the weight of the n^{th} impulse is thus given by $p_n = \int_{(n-1)\Delta}^{n\Delta} p_Y(y)\, dy$. Panel c describes the dithering process; $\{D_n\}$ are independent and identically distributed random variables from a uniform distribution denoted by $p_D(y)$ added to Y'_n to yield Y''_n. **Panel d** is a block diagram modeling the dithering: the input, $p_{Y'}(y)$, passes through a filter whose impulse response is $p_D(y)$. The output of the filter is $p_{Y''}(y)$, the amplitude density of Y''_n. The specific case of $p_D(y)$ given by

$$p_D(y) = \begin{cases} \frac{1}{\delta} & -\frac{\delta}{2} \le y \le \frac{\delta}{2} \\ 0 & \text{elsewhere} \end{cases} \quad .$$

is illustrated.


$\{Y_n\}$
A/D
$\{Y'_n\}$
(a)
$p_Y(y)$
$h_\Delta(y)$
$y = n\Delta$
$p_{Y'}(y)$
$p_n = \int_{(n-1)\Delta}^{n\Delta} p_Y(y)dy$
Δ
....
....
(b)
$\{Y'_n\}$
+
$\{D_n\}$
$\{Y''_n\}$
(c)
$p_{Y'}(y)$
$p_D(y)$
$p_{Y''}(y)$
$\frac{p_n}{\delta}$
δ
....
....
(d)

where $\hat{p}_Y(n\,\Delta)$ is a number between the maximum and minimum of $p_Y(y)$ in the interval $((n-1)\Delta, n\Delta)$. Thus the output of the sampler has the following probability density function

$$p_{Y'}(y) = \sum_i p_i\, \delta_f(y - Y_i)$$

where p_i is as defined in equation (3.15) and $\delta_f(\cdot)$ denotes the impulse function.

As remarked in § 2.3 the definitions of entropy for a random variable in equation (2.3) and for a random vector in equation (2.4) are not valid for a discrete distribution (the integrals in the corresponding equations are not finite for such distributions) and hence the model order determination technique developed cannot be applied to such a time series (i.e., $\{Y'_n\}$). We will assume that the binwidth, Δ was, appropriately chosen to prevent aliasing [12]: Δ must be less than $\frac{1}{2} f_{\max}$ ($f_{\max}$ is the highest frequency component in $p_Y(y)$) to prevent aliasing. The original distribution, $p_Y(y)$, can be recovered from this impulse train by dithering the quantized values (i.e., the output of the A/D converter in figure 3.3a); this dithering is performed by adding an independent random variable to each quantized value. The simplest form of dithering is achieved by adding independent and identically distributed random variables from an uniform distribution, $\{D_n\}$, to the quantized values, $\{Y'_n\}$ (see figure 3.3c); this is essentially convolving the probability density function, $p_{Y'}(y)$ with the uniform probability density function or, equivalently, one could model it as passing the discrete random variable, with probability density function $p_{Y'}(y)$, through a low pass filter (see figure 3.3d); this model is appropriate for determining the probability density function (and hence the entropy) of the dithered random variables. The low pass filter (or the uniform probability density function) is given by

$$p_D(y) = \begin{cases} \frac{1}{\delta} & -\frac{\delta}{2} \le y \le \frac{\delta}{2}. \\ 0 & elsewhere \end{cases}$$

The entropy of the output of this low pass filter will be computed to determine the value of δ that results in this entropy estimate equalling the entropy estimate of the input to the integrator. The output of the low pass filter is given in figure 3.3d; the first order entropy of the output $H_1(Y'')$ is thus given by

$$H_1(Y'') = -\sum_i \delta\, Y''_i \log Y''_i$$

where $Y''_i = \frac{1}{\delta} p_i$; substituting for p_i from equation (3.15) one obtains

$$H_1(Y'') = -\sum_i \Delta\, \hat{p}_Y(i\Delta) \log \left[\frac{\Delta}{\delta}\, \hat{p}_Y(i\Delta) \right].$$

In the limit as $\Delta \rightarrow 0$ and $\delta \rightarrow 0$ and $\frac{\Delta}{\delta} \rightarrow K$ (K is a constant) we have a Riemman integral

$$H_1(Y'') = -\int_{-\infty}^{\infty} p_Y(y) \log p_Y(y)\, dy - \log K.$$

Thus by choosing $\delta = \Delta$ the entropy of Y'' is an unbiased estimator of the true entropy (i.e., the entropy of the input to the intergrator). This theory will be used to estimate the first order entropy in the next section.

3.8. Application: Estimating the model order of spike trains from Cat LSO Units

Of particular interest is the application of entropy methods to estimate the model order of spike trains from the Lateral Superior Olive (LSO). Point Process theory provides a mathematical model for these spike trains [19]; one of the more common approaches is based on the fact that the intensity of the Point process completely characterizes the statistics of the process. In the present context of model order determination the interval between successive events is treated as a time series. It has been shown [7] that the input-output relationship of a system generating inter-event intervals is derivable from the intensity of the desired Point process. It has also been shown [7] that this system is nonlinear for most interesting cases. The nonlinear model described by equation (2.1) is similar to the models that are used to describe the spike trains [8].

The procedure of recording the spike trains from the LSO involves a quantization of the time axis. The low pass filtering, discussed in § 3.7, was provided by small, random perturbations of the values of the time series. The quantization binwidth for the data under consideration was 1 μs; thus the appropriate range for the random perturbations is $[-0.5\times10^{-6}, 0.5\times10^{-6}]$.

The performance of the information-theoretic model order estimator on 13 recordings of spike trains from the LSO is tabulated in Table (3.10); the serial correlation coefficients for these recordings is also included in the table. The largest stationary contiguous portion of the recordings was used in estimating the model order; the details of the procedure used to locate this stationary block are discussed in [8].

Table 3.10

Unit	Stimulus Characteristics		N	ρ	Model Order Predicted
	Freq (kHz)	Level (dB re threshold)			
t108-1c.r2	10.0	19	11776	-0.19	0
t129-1b.r3	15.5	39	12096	-0.24	0
t171-1A.r2	11.5	25	11520	-0.73	1
t176-1E.r2	15.0	30	11072	-0.23	0
t177-1F.r2	15.5	25	12608	-0.33	1
t179-1A.r2	15.5	25	10688	-0.37	1
t182-1C.r11	11.0	33	9472	-0.16	0
t182-1C.r2	11.0	30	12096	-0.29	1
t182-1C.r22	11.0	40	14592	-0.28	0
t184-1A.r6	14.0	18	8192	-0.24	0
t184-1A.r7	14.0	8	8448	-0.16	0
t186-1I.r2	31.5	30	28288	-0.35	1
t186-1I.r3	1.0	0	11456	-0.24	0

(ρ is the serial correlation coefficient; N is the number of inter-spike intervals used in estimating the model order; Freq denotes the stimulus frequency which was equal to the Characteristic frequency of the cell.)

The method of conditional mean [6] estimates the model order of all the LSO data records to be greater than or equal to unity. Thus the Entropy Method is not as sensitive a technique as the conditional mean method. It has been shown [8] that the nonlinear dependence structure of equation (2.1) is appropriate for the LSO data. The performance of the Entropy Method on the LSO data is consistent with its performance on time series with the nonlinear dependence structure under discussion; the results in Table 3.8 indicate that the model order prediction of the Entropy Method is reliable for serial correlation coefficients larger in magnitude than 0.3 which is the trend observed in Table 3.10.

CHAPTER 4

Conclusions

The search for a new method to estimate the model order was primarily motivated by the fact that the existing techniques (i.e., MAICE and MDL) are strongly tied to the assumptions made about the model (linear) and the allowable class of inputs (Gaussian). The Entropy Method does not suffer these limitations as the joint entropy estimate is a distribution-free statistic. However this procedure is insensitive to low dependence levels and is data-intensive. Further the computational resources required are enormous. The fact that the Entropy Method is data-intensive and requires enormous computational resources limits the utility of this technique; it is not feasible to use this technique on time series with large model orders. Thus there is a trade-off in terms of the assumptions made about the model and the performance characteristics of any model order determination technique. This suggests that there probably exists no one "ideal" method that works well on any time series with no assumptions on the model. The Entropy Method should be used in conjunction with other techniques in estimating model order of any given set of observables.

The Entropy Method uses a distribution-free statistic to estimate the joint entropy; this does not imply that the model order determining technique is distribution-free. As was discussed in § 3.4 the conditional entropy, $H(X \mid Y)$ is not very sensitive to the first order coefficient (see figure 3.3) for a Gaussian distribution; this is a limitation imposed by the theory. This theoretical limitation is reflected in the performance of the Entropy Method; the low sensitivity in the performance is comparable to the low sensitivity predicted by the theory. The performance of the Entropy Method on the Exponential AR time series seems to indicate that the theory does not impose such sensitivity constraints for the exponential distribution; this result could not be verified; as discussed in § 2.1, the joint distribution of the elements of a time series is known only in the Gaussian case. The performance on time series with nonlinear dependence lies between the performance on the Gaussian and Exponential AR time series. The low sensitivity observed in the performance of the Entropy Method on time series with nonlinear dependence could be due to theoretical limitations or could be a result of the technique itself.

Bentley [2] suggests an algorithm to determine the nearest neighbour in $\Re^d$ $(d > 1)$; this algorithm has a $O(N_d \log^{d-1} N_d)$ computational complexity as opposed to the brute force approach (see § 3.6) which has a $O(N_d^2)$ complexity. But Bentley himself points out, in a subsequent publication [3], that it is not clear that there is a search structure that implements this algorithm for $d \geq 3$. But if this algorithm could be implemented it would result in considerable savings of computational resources; but for a given time series (i.e., N is fixed) this algorithm could result in savings only to certain dimensions (the critical value of d is given, approximately, by solving $N_d \log^{d-1} N_d = N_d^2$) and for higher dimensions the algorithm is worse than the brute force approach. For example consider $N = 10000$: Bentley's algorithm could result in savings only for $d \leq 6$. Thus even for such a large time series the potential reduction of computational effort is only to the sixth dimension; the situation is worse for smaller time series. In any case it is worth taking a closer look to see if Bentley's algorithm could be implemented.

The manner in which the data points in $\Re^d$ (for $d > 1$) were created (see § 3.1) resulted in the number of elements in $\Re^d$, for a given time series, ($N_d = \left\lfloor \frac{N}{d} \right\rfloor$) decreasing with increasing d; this, as was discussed § 3.5, is one of the reasons responsible for making the Entropy Method data-intensive. The data points in $\Re^d$ (for $d > 1$) can also be created in the following manner: choose all sets of d-adjacent elements in the time series as data points in $\Re^d$. Thus a time series with N elements would yield $N - d + 1$ elements in $\Re^d$. This method of creating data points in $\Re^d$ has the potential to considerably reduce the size of data required for analysis; a detailed study should be made to determine if the Entropy Method, with elements in $\Re^d$ created in the manner just described, is an effective model order estimator.

In § 3.7 a box-car was used to reconstruct the true probability density function of the time series from the given quantized values of the time series; this results in a crude estimate of $p_Y(y)$. One could choose larger values of δ (> Δ) to obtain smoother estimates. One could also obtain smoother estimates of the true density by dithering the quantized data with independent random variables drawn from a triangular distribution, of width 2δ, thereby approximating the true density via linear interpolation . In all these cases one would have to obtain a mathematical expression for the deviation from the entropy of $p_Y(y)$; this may not easy in some cases.

Given a time series, which technique estimates the model order correctly? To answer this question one must be well informed on the strengths and limitations of the different techniques. This stresses the importance of simulations in studying the performance of the different techniques. This performance analysis should cover a variety of dependence structures; the study should include the performance of the technique on varied degrees of dependence. Sensitivity and computational complexity are a couple of performance characteristics that should be studied. For example, a detailed performance analysis of the Entropy Method has been made only on time series with a first order dependence structure; further the study covered only a specific type of nonlinear dependence structure. Thus a careful analysis of a technique should be made to determine its applicability to a specific problem.

References

[1] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol.19, no.6, pp. 716-723, December 1974.

[2] J.L. Bentley and M.I. Shamos, "Divide-and-Conquer in Multidimensional Space," *Proceedings of the 8th Annual ACM Symposium on Theory of Computing*, pp. 220-230, May, 1976.

[3] J.L. Bentley, "Multidimensional Divide-and-Conquer," *Communications of the ACM*, Vol.23, no.4, pp. 214-229, April, 1980.

[4] G.E. Box and G.M. Jenkins, *Time Series Analysis Forecasting and Control.* San Francisco, California: Holden-Day, 1970.

[5] I.S. Gradshteyn and I.M. Ryzhik, *Tables of Integrals, Series and Products.* New York: Academic Press, 1965.

[6] D.H. Johnson, C. Tsuchitani, D.A. Linebarger, and M.J. Johnson, "Application of a Point Process Model to Response of Cat Lateral Superior Olive Units to Ipsilateral Tones," *Hearing Research, to appear.*

[7] D.H. Johnson and D. Linebarger, "Signal Processing Models for Point Processes," *ICASSP Proc.*, 1984.

[8] D.A. Linebarger, "Point Process Models for Discharge Patterns of Single Units in the Lateral Superior Olive of the Cat," M.S. Thesis, Department of Electrical Engineering, Rice University, Houston, Tx, 1984.

[9] E.H. Linfoot, "An Informational Measure of Correlation," *Information and Control*, Vol.1, pp. 85-89, 1957.

[10] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, Vol.63, no.4, pp. 561-580, April 1975.

[11] H. Nakahama, M. Yamamoto, N. Ishii, H. Fujii, and K. Aya, "Dependency as a Measure to Estimate the Order and the Values of Markov Processes," *Biological Cybernetics*, Vol.25, pp. 209-226, 1977.

[12] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing.* New Jersey: Prentice-Hall Inc., 1975.

[13] E. Parzen, *Stochastic Processes.* San Francisco: Holden-Day, Inc., 1962.

[14] K. Pearson, "Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and its Relation to Association and Normal Correlation," in *Karl Pearson's Early Statistical Papers,* E.S. Pearson, Eds. The Cambridge Press: Cambridge, pp. 443-475, 1948.

[15] A. Renyi, "On Measures of Dependence," *Acta Mathematica Academiae Scientiarium Hungaricae*, Vol.10, no.3-4, pp. 441-451, 1959.

[16] H.L. Royden, *Real Analysis.* New York: Macmillan Publishing Co.,Inc., 1963.

[17] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol.6, no.2, pp. 461-464, 1978.

[18] C. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Urbana, Illinois: University of Illinois, 1949.

[19] D.L. Snyder, *Random Point Processes.* New York: Wiley, 1975.

[20] J.B. Thomas, *An Introduction to Statistical Communication Theory.* New York: John Wiley & Sons Inc., 1969.

[21] O. Vasicek, "A Test for Normality Based on Sample Entropy," *Journal Royal Statistical Society*, Vol.38, no.1, pp. 54-59, Series B, 1976.