# NON-PARAMETRIC MAXIMUM ENTROPY

by

Dimitris N. Politis
Purdue University

Technical Report # 91-17

Department of Statistics
Purdue University

April, 1991

△

# NON-PARAMETRIC MAXIMUM ENTROPY

by

Dimitris N. Politis
Purdue University

## ABSTRACT

The standard Maximum Entropy method of Burg (1967) and the resulting autoregressive model has been widely applied for spectrum estimation and prediction. In this article, a generalization of the Maximum Entropy formalism in a non-parametric setting is presented, and the class of the resulting solutions is identified to be a class of Markov processes. The proof is based on a string of information theoretic arguments developed in Choi and Cover's (1984) derivation of Burg's Maximum Entropy spectrum. A framework for the practical implementation of the proposed method is also presented, in the context of both continuous and discrete data.

**Index Terms.** Markov processes, Maximum Entropy, non-linear time series, non-parametric estimation.

## I. Introduction

Suppose $\{X_n, n \in \mathbf{N}\}$ be a wide-sense stationary stochastic process with mean zero and autocovariance $\gamma(k) = EX_tX_{t+k}$, for $k \in \mathbf{Z}$. It is well known (cf. Burg [1], which is reprinted in [3]) that the Maximum Entropy such process that satisfies the constraints $\gamma(i) = c_i, i = 0, 1, \ldots, p$ is the mean zero autoregressive Gaussian process that satisfies these constraints. In fact (cf. Choi and Cover [4]), a wider entropy-maximization problem (where $\{X_n, n \in \mathbf{N}\}$ is *neither* assumed to be wide-sense stationary, *nor* of mean zero) is seen to have the *same* autoregressive Gaussian process as its solution. Note that this linear autoregressive model is actually a special case of a Markov process.

In practical applications, a stretch $X_1, \ldots, X_N$ is observed from the $\{X_n, n \in \mathbf{N}\}$ sequence, from which the autocovariances $\gamma(k)$ for $k = 0, 1, \ldots, p << N$ are estimated accurately. Then, for the purposes of spectral estimation or prediction, the Maximum Entropy principle is invoked, leading to the aforementioned Gaussian autoregressive model. Effectively, the assumption of this model allows for a nontrivial extrapolation of the autocovariance function to places where there are insufficient data, or none at all, i.e. $\gamma(k)$ for $k = p+1, p+2, \cdots, N, N+1, \cdots$.

However, in many situations the observations seem not to be compatible with the Gaussian model to be assumed. Such examples are amply provided by considering any sequence of discrete random variables, e.g. a binary sequence. Another example is the famous Sunspots data-set (monthly means of daily sunspot numbers for the period January 1749 to March 1977), which has been shown to be non-linear and non-Gaussian [15]. Therefore, fitting a linear autoregressive Gaussian model seems to be the wrong thing to do in these cases.

Having discarded the assumption of a Gaussian (and linear) model, attention could be focused on the remaining property of the standard Maximum Entropy solution, namely the Markov property. It would seem desirable to have a non-parametric formulation of the Maximum Entropy principle that permits the extrapolation of distributions, in the same manner Burg's Maximum Entropy extrapolates the autocovariances. This goal is accomplished in the next section, where in fact the solutions to the non-parametric Maximum Entropy problem are shown to be Markov processes.

## II. Markov Processes and Maximum Entropy

Let $\{X_n, n \in \mathbf{N}\}$ be a stochastic process specified by its marginal distribution functions $F_n(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$, $n \in \mathbf{N}$. It will be assumed that, for any $n \in \mathbf{N}$, $F_n(x_1, \ldots, x_n)$ possesses a density $f_n(x_1, \ldots, x_n)$ with respect to a measure $\nu_n$ on $\mathbf{R}^n$. The measure $\nu_n$ will be chosen to be either Lebesgue measure on $\mathbf{R}^n$, or a counting measure, resulting to $f_n(x_1, \ldots, x_n)$'s that are the usual probability density functions (p.d.f.) or probability mass functions (p.m.f.) respectively.

The *entropy* of the $n$-tuple $X_1, \ldots, X_n$ is defined by (cf. [2])

$$H(X_1, \ldots, X_n) = -\int f_n(x_1, \ldots, x_n) \log f_n(x_1, \ldots, x_n) d\nu_n(x_1, \ldots, x_n) \tag{1}$$

Since $H(X_1, \ldots, X_n)$ represents a functional of $f_n$, we can alternatively denote it by $H(f_n)$. Similarly, the *conditional entropy* of $X_n$ given $X_{n-1}, \ldots, X_1$ is defined by

$$H(X_n|X_{n-1}, \ldots, X_1) = -\int f_n(x_1, \ldots, x_n) \log f_n(x_n|x_{n-1}, \ldots, x_1) d\nu_n(x_1, \ldots, x_n) \tag{2}$$

where $f_n(x_n|x_{n-1}, \ldots, x_1)$ is the conditional density of $X_n$ given the values of $X_{n-1}, \ldots, X_1$.

The stochastic process $\{X_n, n \in \mathbf{N}\}$ is said to have an *entropy rate*

$$h_X = \lim_{n \to \infty} \frac{H(X_1, \ldots, X_n)}{n} \tag{3}$$

provided the limit exists. It can be shown that the limit exists for (strictly) stationary processes.

The following lemma will be of use in proving our main theorem. Its essence is that extrapolation of distributions is possible in the class of stationary Markov processes.

**Lemma 1** *Let $\{Z_n, n \in \mathbf{N}\}$ be a $(p-1)$th order strictly stationary Markov process, where $p$ is a positive integer. All the marginal densities of the process $\{Z_n, n \in \mathbf{N}\}$ are completely determined by its pth order marginal density $g(z_1, \ldots, z_p)$.*

**Proof.** Note that by the term 0-order Markov process, it is meant an i.i.d. (independent and identically distributed) sequence. Hence for $p = 1$ the lemma is obvious, and for any $n \in \mathbf{N}$ the $n$th order marginal density (i.e. the density of $Z_1, \ldots, Z_n$) is $g(z_1, \ldots, z_n) = g(z_1)g(z_2) \cdots g(z_n)$.

Now let $p > 1$. Note that the $p$th order marginal density $g(z_1, \ldots, z_p)$ can be expressed as

$$g(z_1, \ldots, z_p) = g(z_p | z_{p-1}, \ldots, z_1) g(z_1, \ldots, z_{p-1}) \tag{4}$$

where $g(z_k | z_{k-1}, \ldots, z_1)$ is the conditional density of $Z_k$ given $Z_{k-1}, \ldots, Z_1$ (which, by stationarity, is identical to the conditional density of $Z_{k+n}$ given $Z_{k+n-1}, \ldots, Z_{n+1}$, for any $n \in \mathbf{N}$). From equation (4), the functional form of $g(z_p | z_{p-1}, \ldots, z_1)$ (and therefore also of $g(z_k | z_{k-1}, \ldots, z_{k-p+1}), \forall k \geq p$) can be obtained. Then, for any $n \geq p$, the $n$th order marginal density is given by the chain rule

$$g(z_1, \ldots, z_n) = g(z_1, \ldots, z_{p-1}) \prod_{k=p}^{n} g(z_k | z_{k-1}, \ldots, z_1) = g(z_1, \ldots, z_{p-1}) \prod_{k=p}^{n} g(z_k | z_{k-1}, \ldots, z_{k-p+1}) \tag{5}$$

where the Markov property was explicitly used.□

Let $g(z_1, \ldots, z_p)$ be a density with respect to measure $\nu_p$ on $\mathbf{R}^p$. We will say that $g(z_1, \ldots, z_p)$ satisfies the *stationarity requirement* if there exists in some probability space a sequence of random variables $Z_1, \ldots, Z_p$ with (joint) density $g(z_1, \ldots, z_p)$ that is stationary, i.e., for any $k = 1, 2, \ldots, p$, the joint distribution of $Z_1, \ldots, Z_k$ is identical to the joint distribution of $Z_{1+n}, \ldots, Z_{k+n}$, for any $n = 1, 2, \ldots, p-k$. This is equivalent to requiring that the marginals of $g(Z_1, \ldots, Z_p)$ are 'right', in the sense that they are compatible with the hypothesis of stationarity.

**Lemma 2** *For some positive integer $p$, let $g(x_1, \ldots, x_p)$ be a density with respect to measure $\nu_p$ on $\mathbf{R}^p$ that satisfies the stationarity requirement. Then, there exists a $(p-1)$th order strictly stationary Markov process $\{Z_n, n \in \mathbf{N}\}$ with pth order marginal density equal to $g(z_1, \ldots, z_p)$.*

**Proof.** For $p = 1$ the lemma is obviously true. So assume $p > 1$. Let $Z_1, \ldots, Z_p$ have (joint) density $g(z_1, \ldots, z_p)$. For $n > p$, construct the $Z_n$'s inductively, by letting the conditional density of $Z_n$ given that $Z_{n-1} = z_{n-1}, \ldots, Z_1 = z_1$ be equal to

$$g(z_n | z_{n-1}, \ldots, z_{n-p+1}) \equiv \frac{g(z_{n-p+1}, \ldots, z_n)}{g(z_{n-p+1}, \ldots, z_{n-1})}$$

From the construction it is apparent that the sequence $\{Z_n, n \in \mathbf{N}\}$ is $(p-1)$th order Markov, with time-invariant transition probability density function $g(z_n | z_{n-1}, \ldots, z_{n-p+1})$, i.e.

a function whose functional form does not depend on $n$. To complete the proof, it has to be shown that $\{Z_n, n \in \mathbf{N}\}$ is also stationary. It would then be implied that it possesses a $p$th order marginal density equal to $g(z_1, \ldots, z_p)$.

In order to do this, consider the process $\{Y_n, n \in \mathbf{N}\}$, where, for any integer $n$, $Y_n$ is defined to be the block of $(p-1)$ consecutive observations from the $Z$-sequence starting from $Z_n$. For example, $Y_1 = (Z_1, \ldots, Z_{p-1}), Y_2 = (Z_2, \ldots, Z_p)$, and so forth.

It is apparent that the $\{Y_n\}$ sequence is first-order Markov, and it is stationary if and only if the $\{Z_n\}$ sequence is stationary. It is also easy to see that the conditional density of $Y_2$ given $Y_1$ *coincides* with the conditional density of $Z_p$ given $Z_{p-1}, \ldots, Z_1$, using the appropriate notation, and the marginal density of $Y_1$ *coincides* with the marginal density of $Z_1, \ldots, Z_{p-1}$. Now it is immediate that the (unconditional) density of $Y_2$ coincides with that of $Y_1$, showing that the sequence $\{Y_n\}$ is stationary.□

The point to be made from Lemma 2 is the following. Suppose that a density $g(x_1, \ldots, x_p)$ is given, with the additional knowledge that $g(x_1, \ldots, x_p)$ is the $p$th order marginal density of some stationary process. Among the many stationary processes that have $p$th order marginal density equal to $g(x_1, \ldots, x_p)$, there is one that is actually $(p-1)$th order Markov. As a matter of fact, this Markov process can also be characterized as a maximum entropy process.

**Theorem 1** *For some positive integer p, let $g(x_1, \ldots, x_p)$ be a density with respect to measure $\nu_p$ on $\mathbf{R}^p$ that satisfies the stationarity requirement. The strictly stationary stochastic process $\{X_n, n \in \mathbf{N}\}$ that has maximum entropy rate subject to the constraint*

$$f_p(x_1, \ldots, x_p) = g(x_1, \ldots, x_p) \tag{6}$$

*for all $(x_1, \ldots, x_p) \in \mathbf{R}^p$, is the $(p-1)$th order (strictly) stationary Markov process satisfying the constraint.*

**Proof.** The proof is based on the same arguments used in the information theoretic proof of Burg's Maximum Entropy spectrum presented in Choi and Cover [4]. Assume that the observations $X_1, \ldots, X_p$ from the stationary process $\{X_n, n \in \mathbf{N}\}$ satisfy the constraint (6).

First consider the case $p = 1$. If we let $\{Z_n, n \in \mathbf{N}\}$ be an i.i.d. sequence, with the density of $Z_1$ given by $g(z_1)$, then we have

$$H(X_1,\ldots,X_n) \leq \sum_{k=1}^{n} H(X_k) = \sum_{k=1}^{n} H(Z_k) = H(Z_1,\ldots,Z_n) \tag{7}$$

Consider now the case $p > 1$, and let $\{Z_n, n \in \mathbf{N}\}$ be a $(p-1)$th order (strictly) stationary Markov process, whose $n$th dimensional marginal density $g(z_1,\ldots,z_n)$ is given by

$$g(z_1,\ldots,z_n) = g(z_1,\ldots,z_{p-1}) \prod_{k=p}^{n} g(z_k|z_{k-1},\ldots,z_{k-p+1}) \tag{8}$$

from Lemma 2, for any $n \geq p$. Then we have

$$\begin{aligned} H(X_1,\ldots,X_n) &\overset{(a)}{=} H(X_1,\ldots,X_p) + \sum_{k=p}^{n} H(X_k|X_{k-1},\ldots,X_1) \\ &\overset{(b)}{\leq} H(X_1,\ldots,X_p) + \sum_{k=p}^{n} H(X_k|X_{k-1},\ldots,X_{k-p+1}) \\ &\overset{(c)}{=} H(Z_1,\ldots,Z_p) + \sum_{k=p}^{n} H(Z_k|Z_{k-1},\ldots,Z_{k-p+1}) \overset{(d)}{=} H(Z_1,\ldots,Z_n) \end{aligned} \tag{9}$$

Here, (a) is the chain rule for entropy; (b) follows from the conditional entropy inequality $h(A|B,C) \leq h(A|B)$; (c) is consequence of the fact that the conditional entropies involved are uniquely determined by functionals of the $p$th order marginal density, which is the same for both $\{X_n\}$ and $\{Z_n\}$ processes; and (d) follows from the chain rule for entropy in connection with the Markov structure of $\{Z_n, n \in \mathbf{N}\}$.

Note that the limit $h_Z = \lim_{n\to\infty} \frac{H(Z_1,\ldots,Z_n)}{n}$ can be explicitly calculated yielding

$$h_Z = \begin{cases} -\int g(z_1)\log g(z_1) d\nu_1(z_1) & \text{if } p = 1 \\ -\int g(z_1,\ldots,z_p)\log g(z_p|z_{p-1},\ldots,z_1) d\nu_p(z_1,\ldots,z_p) & \text{if } p > 1 \end{cases} \tag{10}$$

Combining results (7) and (9), it is seen that $H(X_1,\ldots,X_n) \leq H(Z_1,\ldots,Z_n)$, for any value of $p \in \mathbf{N}$. By stationarity, is follows that the limit $h_X = \lim_{n\to\infty} \frac{H(X_1,\ldots,X_n)}{n}$ exists as well. It is then immediate that $h_X \leq h_Z$, and the theorem is proved. □

It is apparent from the proof that the only occasion constraint (6) was invoked, was in step (c), and only through a functional of $f_p(x_1,\ldots,x_p)$ and $g(x_1,\ldots,x_p)$ that remains unchanged if $f_p$ and $g$ are different on a set of $\nu_p$-measure zero. Hence, the theorem remains true if the constraint (6) is substituted by $f_p(x_1,\ldots,x_p) = g(x_1,\ldots,x_p)$, almost everywhere with respect to $\nu_p$.

Of course, the distribution functions corresponding to densities coinciding almost everywhere with respect to the dominating measure are identical. In this sense, we can say that the Markov process $\{Z_n\}$ in Lemma 2 and Theorem 1 is *uniquely* specified, meaning that its marginal distribution functions are completely determined. However, there are other types of constraints different from (6) subject to which the Maximum Entropy problem is well defined and has a uniquely determined solution.

One such example is Burg's problem, where the constraint consists of a restriction on the second order moment structure of the $p$th marginal. Another example is provided by a result of Shannon ([14], Appendix 4) concerning the capacity of communication channels, that has also found applications in ergodic theory [10]. To briefly describe it, suppose that $\{X_n, n \in \mathbf{N}\}$ is a strictly stationary process, with $X_1$ taking values in the finite set $\{a_1, \ldots, a_k\}$. Let $B = (b_{ij}), i, j = 1, \ldots, k$ be an irreducible and aperiodic matrix of zeroes and ones indicating the permissible transitions of the sequence, i.e. $b_{ij} = 1$ if $P(X_{n+1} = a_j | X_n = a_i) > 0$, and $b_{ij} = 0$ if $P(X_{n+1} = a_j | X_n = a_i) = 0$. Then, the process that has Maximum Entropy among such processes is a first order Markov chain with transition probabilities that can be explicitly calculated.

Hence in general the constraint might involve only a certain feature of the $p$th order marginal distribution of $\{X_n, n \in \mathbf{N}\}$. It is the purpose of the next theorem to show that, whatever the type of constraint, maximization of entropy leads to a $(p-1)$th order Markov process, provided of course the problem is well defined.

**Theorem 2** *Let $\mathcal{F}_p$ be the set of all probability distribution functions on $\mathbf{R}^p$, and let $T : \mathcal{F}_p \to \mathcal{T}$ be some functional with range $\mathcal{T}$, where $\mathcal{T}$ is some space. Let $\tau$ be a subset of $\mathcal{T}$, and denote by $\mathcal{S}_\tau$ the set of all strictly stationary processes with pth order marginal distribution $F_p$ satisfying $T(F_p) \in \tau$. Denote by $\mathcal{M}_\tau$ the intersection of $\mathcal{S}_\tau$ with the set of all $(p-1)$th order Markov processes. Then*

$$h(\mathcal{S}_\tau) \leq h(\mathcal{M}_\tau) \tag{11}$$

*where $h(\mathcal{S}) \equiv \sup\{h_X : \{X_n\} \in \mathcal{S}\}$ is the supremum of the entropy rate $h_X$, with the process $\{X_n\}$ ranging over the class $\mathcal{S}$.*

**Proof.** First note that $\mathcal{M}_\tau$ is empty if and only if $\mathcal{S}_\tau$ is empty too, and hence equation (11) would hold by defining the supremum of the empty set to be negative infinity.

Let the process $\{X_n\}$ be in $\mathcal{S}_\tau$, and let $\{Z_n\}$ be in $\mathcal{M}_\tau$. The process $\{Z_n\}$ can be constructed by extrapolation of the $p$th marginal distribution of $\{X_n\}$ analogously to the results of Lemma 1 and Lemma 2. Now by a similar chain of arguments as in (9) it is shown that $h_X \leq h_Z$, and the theorem is proved. □

## III. Non-Parametric Maximum Entropy

The success of Burg's Maximum Entropy method is both due to its intuitive appeal, as well as its easy implementation, which is in essence fitting a parametric (Gaussian autoregressive) model to the data. In this section, a procedure for the Non-Parametric implementation of the Maximum Entropy Method based on Theorem 1 will be described that is equally straightforward.

Suppose that a stretch $X_1, \ldots, X_N$ is observed from the *strictly stationary* process $\{X_n, n \in \mathrm{N}\}$. An integer $p << N$ is then chosen, and the $p$th order marginal density $g(x_1, \ldots, x_p)$ is estimated from the data in a standard non-parametric fashion. This would involve calculating the observed relative frequencies in the case of discrete random variables, or the usual kernel-smoothed multivariate density estimates for (absolutely) continuous random variables [13]. Then, for the purposes of extrapolation of distributions or prediction, the Maximum Entropy Principle is invoked, implying (in connection with Theorem 1) that the distribution of the data should be approximated by the distribution of a $(p-1)$th order stationary Markov process possessing a $p$th order marginal density equal to the estimated one. The problem of choosing $p$ properly will be separately addressed in Section V, since it is of importance in practice.

Note that some condition of weak dependence in the sequence $\{X_n\}$ must be assumed in order to have consistent density estimates, (and to be able to reasonably approximate $\{X_n\}$ by a Markov model). For example, it may be assumed that the $\{X_n\}$ sequence satisfies some mixing condition (cf. [6] and the references therein). In connection with real-valued Markov processes, the most common weak dependence condition is $\phi$-mixing, which is equivalent in this case to assuming that the sequence of $\phi$-mixing coefficients is decreasing exponentially fast [6]. For countable-state Markov processes, the $\alpha$-mixing condition should be sufficient, corresponding to the Markov process being aperiodic and ergodic [11].

However, some special care should be taken in order for our estimated density to satisfy the stationarity requirement. In fact, the estimated marginal density $\hat{f}_p(x_1, \ldots, x_p)$ will typically *not* satisfy the stationarity requirement, as the following simple example shows. Suppose $\{X_n\}$ is a 0–1 binary sequence from which the sample $X_1 = 0, X_2 = 1$ is observed. It is obvious

that the estimated second order marginal density (p.m.f.) which puts mass one on the point $(0,1) \in \{0,1\}^2$ is not compatible with the hypothesis of stationarity. Although it can be argued that, for large sample sizes $N$, the estimated marginal density will be 'close' to satisfying the stationarity requirement, it seems difficult to assess and quantify this 'closeness'.

The way out of this difficulty is once again presented in the framework of Markov processes. For a $(p-1)$th order Markov process, estimation of the $p$th order marginal density $g(x_1, \ldots, x_p)$ is equivalent to simultaneously estimating $g(x_1, \ldots, x_{p-1})$ together with the conditional density $g(x_p|x_{p-1}, \ldots, x_1)$. The crucial observation now is that if the Markov process is assumed to be stationary, the conditional density $g(x_p|x_{p-1}, \ldots, x_1)$ *determines uniquely* the marginal $g(x_1, \ldots, x_{p-1})$ under some conditions. As a matter of fact, $g(x_1, \ldots, x_{p-1})$ would be the so-called *stationary* or *invariant* marginal of the Markov process. The implication then is that *only the conditional density* $g(x_p|x_{p-1}, \ldots, x_1)$ *should be estimated from the data, and* $g(x_1, \ldots, x_{p-1})$ *should be set to be the corresponding stationary marginal.*

To elaborate, let $\{Z_n, n \in \mathbf{N}\}$ be a $(p-1)$th order stationary Markov process. Then it is immediate that the process $\{Y_n, n \in \mathbf{N}\}$, which was constructed in the proof of Lemma 2, is first-order Markov and stationary. As it was mentioned, the conditional density of $Y_2$ given $Y_1$ coincides with the conditional density of $Z_p$ given $Z_{p-1}, \ldots, Z_1$, using the appropriate notation, and the stationary marginal of $Y_1$ coincides with the stationary marginal density of $Z_1, \ldots, Z_{p-1}$.

Now suppose that $\{Y_n\}$ takes values in the countable set $\{a_1, a_2, \ldots\}$. If the matrix $B = (b_{ij})$ of transition probabilities $b_{ij} = P(Y_2 = a_j | Y_1 = a_i)$ is positive recurrent, irreducible and aperiodic, it is well-known [12] that there is a unique stationary marginal probability distribution $P^\star$ for $Y_1$, given by any of the rows of the matrix $B^\star = \lim_{n\to\infty} B^n$. As a result of positive recurrence, $P^\star(Y_1 = a_j) > 0$, for any $j \in \mathbf{N}$. Also note that if $\{Y_n\}$ takes values in a finite set, the matrix $B$ will necessarily be positive recurrent. Hence, in this case, the stationary marginal $g(z_1, \ldots, z_{p-1})$, (which is identical to the stationary marginal of $Y_1$) is uniquely determined by the conditional density $g(z_p|z_{p-1}, \ldots, z_1)$, (which is identical to the conditional density of $Y_2$ given $Y_1$).

In case $\{Y_n\}$ takes values in an uncountable set, e.g. a possibly infinite rectangle in $\mathbf{R}^{p-1}$,

(that corresponds to the $\{Z_n\}$ process being real-valued), then some additional regularity conditions must be imposed to ensure that the conditional density $g(z_p|z_{p-1},\ldots,z_1)$ determines uniquely the stationary marginal $g(z_1,\ldots,z_{p-1})$. The usual assumption in this connection is that $\{Y_n\}$ satisfies Doeblin's condition [5], that actually turns out to be equivalent to an exponentially decreasing sequence of $\phi$-mixing coefficients [11].

It is apparent by the above discussion that, in a number of interesting cases, the stationary marginal $g(z_1,\ldots,z_{p-1})$ of a $(p-1)$th order stationary Markov process $\{Z_n\}$ is uniquely determined by the conditional density $g(z_p|z_{p-1},\ldots,z_1)$. This observation justifies the formulation of the following theorem, which is tailor-made for the implementation of the proposed Non-Parametric Maximum Entropy Method.

**Theorem 3** *For some integer $p>1$, let $\{Z_n^\star, n\in\mathbf{N}\}$ be a $(p-1)$th order stationary Markov process, whose conditional density $g(z_p|z_{p-1},\ldots,z_1)$ uniquely determines $g(z_1,\ldots,z_{p-1})$, the stationary marginal density with respect to measure $\nu_p$ on $\mathbf{R}^p$. Then $\{Z_n^\star\}$ has maximum entropy rate in the class of all strictly stationary processes $\{X_n, n\in\mathbf{N}\}$ whose conditional density of $X_p$ given $X_{p-1},\ldots,X_1$, satisfies*

$$f_p(x_p|x_{p-1},\ldots,x_1)=g(x_p|x_{p-1},\ldots,x_1) \tag{12}$$

*for almost all points $(x_1,\ldots,x_p)$ with respect to $\nu_p$.*

**Proof.** Let $\{X_n, n\in\mathbf{N}\}$ be a strictly stationary process with $n$-dimensional marginal density $f_n(x_1,\ldots,x_n), n\in\mathbf{N}$, that satisfies (12). This implies that the $p$th marginal density of $\{X_n\}$ is given by $f_p(x_1,\ldots,x_p)=f_{p-1}(x_1,\ldots,x_{p-1})g(x_p|x_{p-1},\ldots,x_1)$.

By Lemma 2, a $(p-1)$th order (strictly) stationary Markov process $\{Z_n, n\in\mathbf{N}\}$ can be constructed, so as to have $p$th order marginal density equal to $f_p(x_1,\ldots,x_p)$. Then, from Theorem 1 it follows that $h_X\le h_Z$.

However, it is obvious that the process $\{Z_n\}$ also satisfies (12). But from the assumptions of the theorem it follows that there is a unique $(p-1)$th order stationary Markov process satisfying (12), and this is $\{Z_n^\star\}$. Hence, the processes $\{Z_n\}$ and $\{Z_n^\star\}$ should coincide, and the theorem is proved.□

## IV. Extrapolation of Distributions and Prediction

As shown in Section II, the solutions to the Non-Parametric Maximum Entropy problem turn out to be Markov processes. Considering that the class of Markov processes is the 'natural' class in which extrapolation of distributions is possible (cf. Lemma 1), the ability to extrapolate seems to be intimately linked with the Maximum Entropy Principle.

Burg's Maximum Entropy Method can also be viewed from that angle, that is, as a method for distributional extrapolation in the Gaussian setting. Since Gaussian processes are completely determined by their second order moment structure, the extrapolation of distributions is equivalent in this case to the extrapolation of autocovariances and spectral estimation.

An additional feature of Markov processes (and hence processes with Maximum Entropy) is that they provide an ideal framework for the purposes of prediction of $X_{N+1}$ given the values of $X_N, \ldots, X_1$. The autoregressive model of Burg yields the obvious predictor, linear in $X_N, \ldots, X_{N-p+1}$, which is optimal with respect to Mean Squared Error in the Gaussian case.

In our general non-parametric setting, the function $q(x_N, \ldots, x_1)$ that minimizes the Mean Squared Error of prediction $E(X_{N+1} - q(X_N, \ldots, X_1))^2$, is the conditional expectation [9] $q(x_N, \ldots, x_1) = E(X_{N+1}|X_N = x_N, \ldots, X_{N-p+1} = x_{N-p+1})$, which is computable (knowing the conditional density $f_p(x_p|x_{p-1}, \ldots, x_1)$), and does not depend on the values of $X_1, \ldots, X_{N-p}$ (by the Markov property).

Of course, in the practical problem the conditional density $f_p(x_p|x_{p-1}, \ldots, x_1)$ is not exactly known. Nevertheless, an estimate $\hat{f}_p(x_p|x_{p-1}, \ldots, x_1)$ can be constructed based on the data, as outlined in the previous section. Using the estimated (rather than the exact) $p$th order density corresponds to calculating the prediction function

$$\hat{q}_{N,p}(x_N, \ldots, x_1) = \int x_{N+1} \hat{f}_p(x_{N+1}|x_N, \ldots, x_{N-p+1}) d\nu_1(x_{N+1}) \tag{13}$$

which is asymptotically optimal with respect to Mean Squared Error (cf. [6]). In the Gaussian case, the same argument would apply to justify the use of the linear predictor with estimated coefficients in the autoregressive model of Burg's method.

## V. Choosing the Order of the Model

Let us now return to the problem of choosing $p$ properly. In the case of fitting an autoregressive model (as in Burg's method), choosing $p$ amounts to choosing the order of the autoregression. Many data-driven criteria have been formulated for the choice of $p$ in this case, the most popular ones being Akaike's information criterion (AIC) and Bayesian information criterion (BIC).

Note that in the autoregressive case, the number of parameters to be estimated from the data is linear in $p$. However, in the non-parametric setting under consideration, the number of parameters to be estimated from the data is exponential in $p$ for finite-state Markov chains, or infinite otherwise. This observation is sufficient to rule out minimization of AIC or BIC as a sensible way of choosing the order $p$.

Nevertheless, an attractive alternative criterion for a data-driven choice of $p$ exists, in the form of ordinary or predictive cross-validation. Actually, it turns out that, applied to autoregressive model fitting, predictive cross-validation is approximately equivalent to minimizing the BIC, and ordinary cross-validation is approximately equivalent to minimizing the AIC [8]. In this sense, ordinary and predictive cross-validation can be considered as the extensions of the AIC and BIC criteria respectively in more general contexts.

In brief, the proposed cross-validation procedures would go as follows. As in the previous section, with the order of the model chosen to be $p$, the (approximately) optimal predictor of $X_{N+1}$ given that $X_N = x_N, \ldots, X_1 = x_1$ would be $\hat{q}_{N,p}(x_N, \ldots, x_1)$, as given in equation (13). Define the usual residuals $\epsilon_{N,p}(s)$ by

$$\epsilon_{N,p}(s) = X_s - \hat{q}_{N,p}(X_{s-1}, \ldots, X_1) \tag{14}$$

for $s = 1, \ldots, N$, and using $X_k = 0$, for $k \leq 0$. For a given data-set, the sum of squares $E_{N,p} = \frac{1}{N}\sum_{s=1}^{N} \epsilon_{N,p}^2(s)$ can be computed for $p = 1, 2, \cdots$. Minimizing $E_{N,p}$ with respect to $p$ constitutes the (ordinary) cross-validation method of choosing $p$.

However, in order to find the order of the model that has 'best' performance as measured by the accuracy of its predictions, it seems pertinent to look at the predictive (or recursive) residuals $e_p(s)$ that are defined in a slightly different way. Recall that for the computation

of the usual residuals $\epsilon_{N,p}(s)$, the predictor $\hat{q}_{N,p}$ was used, which was based on the *whole* sample $X_1, \ldots, X_N$. Define $\bar{q}_{s+1,p}(X_s, \ldots, X_1)$, for $s = 0, 1, \ldots$, to be the (approximately) optimal predictor of $X_{s+1}$ given that $X_s = x_s, \ldots, X_1 = x_1$. Note that for estimation of $\bar{q}_{s+1,p}(X_s, \ldots, X_1)$ *only* the sample $X_1, \ldots, X_s$ is to be used, (and the assumption $X_k = 0$, for $k \leq 0$). In this sense, the predictive residuals $e_p(s) = X_s - \bar{q}_{s,p}(X_{s-1}, \ldots, X_1)$, for $s = 1, \ldots, N$, provide a measure of the accuracy of predictions using a model of order $p$. Minimizing $\bar{E}_{N,p} = \frac{1}{N}\sum_{s=1}^{N} e_p^2(s)$ with respect to $p$ constitutes the predictive cross-validation method of choosing the order $p$.

Carrying through the predictive cross-validation procedure involves an increased computational effort as compared to using ordinary cross-validation, mainly because, for each value of $p$, $N$ different one-step predictors $\bar{q}_{s,p}, s = 1, \ldots, N$, have to be calculated. That this extra effort might be worthwhile is exemplified by the familiar case of a linear autoregressive model of true order $p^\star$. Using the BIC criterion (or predictive cross-validation) provides an asymptotically consistent means to estimate the true order of the model from the data, while it is well known that use of the AIC criterion (or ordinary cross-validation) does not necessarily lead to a consistent estimate. Comparing the performance of ordinary and predictive cross-validation in our more general non-parametric setting shall be the subject of further research.

As a final note, there is a lot to be said in favor of using *finite-state* Markov models as an approximation, even if the time series under investigation is real-valued. To elaborate, suppose that $\{X_n, n \in \mathbf{N}\}$ takes values in the finite set $\{a_1, \ldots, a_k\}$. Then, in order to estimate the $p$th order marginal density and approximate $\{X_n\}$ by a $(p-1)$th order Markov process $\{Z_n\}$ with the same $p$th marginal, about $k^p$ parameters should be estimated from the data $X_1, \ldots, X_N$. Indeed, this problem can be compared to estimating the probabilities in a multinomial model with $k^p$ cells. It is intuitively clear that to have reasonable estimates (and few empty cells when counting relative frequencies) it must be the case that $k^p << N$, and rather $k^p < \sqrt{N}$. So, as 'a rule of thumb', one would say that $p$ should be taken to be of smaller order than $\frac{1}{2}\log_k N$ in this case. The problem is that for other than finite-state models, the number of parameters to be estimated from the data is infinite, and our intuition breaks down.

For countable-state (say integer-valued) processes, *truncation* seems to be the obvious prac-

tical approximation, i.e. to artificially reduce the state-space from $\{1, 2, \cdots\}$ to $\{1, 2, \cdots, k, k^+\}$, where the symbol $k^+$ corresponds to observations in the set $\{k+1, k+2, \cdots\}$. For uncountable-state (say real-valued) processes, the practical problems are more serious. To start with, due to the sparsity of $p$-dimensional Euclidean space, the rate of convergence of the multivariate density estimate $\hat{f}_p(x_1, \ldots, x_p)$ is very slow [13]. This implies that the estimate of the conditional density

$$\hat{f}_p(x_p|x_{p-1}, \ldots, x_1) \equiv \frac{\hat{f}_p(x_1, \ldots, x_p)}{\hat{f}_{p-1}(x_1, \ldots, x_{p-1})} \tag{15}$$

also converges very slowly. In addition, the assumed $\phi$-mixing (or Doeblin's) condition is a very strong form of weak dependence that is satisfied only rarely. For example, in the case of Gaussian processes the $\phi$-mixing condition is equivalent to $m$-dependence [7]. Hence the *only* Caussian $\phi$-mixing Markov process is an i.i.d. (independent and identically distributed) sequence of normal random variables. The practical solution would be to artificially discretize the state-space by dividing the real line into cells of appropriate sizes, and to consider the resulting finite-state model. This procedure is closely related to using *histograms* instead of kernel-smoothed functions for density estimation.

## VI. Conclusions

The stationary processes with maximum entropy rate in the class of processes whose $p$th order marginal distribution satisfies some constraint were shown to be Markov processes. In particular, the $(p-1)$th order stationary Markov process with $p$th order marginal density $g(z_1, \ldots, z_p)$, with respect to some measure, was shown to possess maximum entropy rate in the class of stationary processes with $p$th order marginal density equal to $g(z_1, \ldots, z_p)$. This result forms the basis for an extension of the usual Gaussian Maximum Entropy Method of Burg to non-parametric settings.

A framework for the practical implementation of the proposed Non-Parametric Maximum Entropy Method was also presented. Specifically, the $p$th order marginal density should be estimated from the observed data, taking care that the estimate is compatible with the hypothesis of stationarity. This can be achieved by estimating the conditional density $g(x_p|x_{p-1}, \ldots, x_1)$ from the data, and setting $g(x_1, \ldots, x_{p-1})$ to be the corresponding stationary marginal. Then, for the purposes of extrapolation of distributions or prediction, the Maximum Entropy Principle can be invoked, implying that the distribution of the data can be approximated by that of a $(p-1)$th order stationary Markov process with the estimated $p$th order marginal density. Notably, unless the estimated $p$th order density is multivariate Gaussian, the Non-Parametric Maximum Entropy Method would point to a *non*-linear Markov model. Finally, the important problem of choosing the order $p$ of the fitted model was addressed, and the cross-validation methodology was suggested as its possible solution.

## VII. Acknowledgement

The author is grateful to Tom Cover, D. Gatzouras, Steve Lalley, and Herman Rubin, for many helpful discussions.

## References


[1] Burg, J.P. (1967), Maximum Entropy Spectral Analysis, *Proc. 37th Ann. Int. Mtg. Soc. Explor. Geophys.*, Oklahoma City.

[2] Cover, T.M. and Thomas, J. (1991), *Elements of Information Theory*, John Wiley, New York.

[3] Childers, D. G. (1978), *Modern Spectrum Analysis*, IEEE Press, New York.

[4] Choi, B.S. and Cover, T.M. (1984), An Information-Theoretic Proof of Burg's Maximum Entropy Spectrum, *Proc. IEEE*, vol. 72, 8, 1094-1095.

[5] Doob, J.L. (1953), *Stochastic Processes*, John Wiley, New York.

[6] Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989), *Nonparametric Curve Estimation from Time Series*, Lecture Notes in Statistics No.60, Springer-Verlag.

[7] Ibragimov, I.A. and Linnik, Y.V. (1971), *Independent and stationary sequences of random variables*, Wolters-Noordhoff, Groningen.

[8] Kavalieris, L. (1989), The Estimation of the Order of an Autoregression Using Recursive Residual and Cross-Validation, *J. Time Ser. Anal.*, vol.10, No.3, 271-282.

[9] Papoulis, A. (1984), *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York.

[10] Parry, W and Tuncel, S. (1982), *Classification Problems in Ergodic Theory*, Cambridge University Press.

[11] Rosenblatt, M. (1971), *Markov Processes. Structure and Asymptotic Behavior*, Springer-Verlag.

[12] Ross, S.M. (1983), *Stochastic Processes*, John Wiley, New York.

[13] Ruschendorf, L. (1977), Consistency of estimators for multivariate density functions and for the mode, *Sankhya, Ser. A*, 39, 243-250.

[14] Shannon, C.E. and Weaver, W. (1963), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.

[15] Subba Rao, T. and Gabr, M.M. (1980), A test for linearity of stationary time series, *J. Time Ser. Anal.*, 1, 145-158.