Division of Research
Graduate School of Business Administration
The University of Michigan

September 1981

TIME SERIES ANALYSIS OF BINARY DATA

Working Paper No. 277

Daniel McRae Keenan

The University of Michigan

FOR DISCUSSION PURPOSES ONLY

Acknowledgement.

This research was funded by a grant from the National Science Foundation, No. MCS76-81435. This work was part of the author's Ph.D. thesis submitted to the Department of Statistics, University of Chicago. The author is grateful to Professor Ronald A. Thisted under whose guidance this work was carried out.

1. Introduction

In both experimental and nonexperimental settings measurements taken sequentially in time have become quite common. For example, we may be observing the same machine or individual over time; the position of a satellite and stress level measurements of an individual are data of this form. The theory for continuous-valued random variables is rather well developed: e.g., regression analysis, time series analysis, etc.; however, for discrete-valued data, different approaches are required. For independent observations, logit, probit, and log-linear models are available. This is the domain of qualitative data analysis. When the observations are dependent and discrete-valued, stochastic process models are relevant. Markov chain theory is well developed; Billingsley (1961) and Kemeny and Snell (1960) are relevant references for statistical inference on Markov chains. However, problems often exhibit more structure than just stationary transition probabilities. This work is concerned with developing time series models for discrete-valued data, allowing for arbitrarily long memory. The learning theory models and chains of infinite-order are related, but are concerned with different structure (Lamperti and Suppes (1959), Bush and Mosteller (1951)).

The data in the following three examples typify a certain type of data; the observations are sequential in time, discrete-valued, and correlated. In labor economics an important problem is the determination of those factors which influence labor force participation. The Bureau of Labor Statistics records over time (e.g., monthly) whether various individuals are in the labor force or not (Heckman (1979)). In psychology one can study the moods of individuals over time in order to determine the variability both within and across individuals (Larson (1979)). For example, if the happy-sad continuum (i.e., semantic differential) is divided into two parts, happy and sad, then the data would consist of daily recordings of zeroes and ones. In business the

prediction of directional changes of business cycles is of concern. If the data consists of only the past directional changes, then the data is a series of zeroes and ones. In these three examples the data, $(D_1, D_2, \ldots D_n)$, was discrete-valued (actually, binary), sequential in time and correlated.

The approach of this paper is quite general; it is assumed that an underlying time series of continuous-valued data generates the time series of discrete-valued data. The family of discretization mechanisms is broad and can often be given intuitive meaning. The underlying probability structure is exploited with much of the structure carrying over to the discrete process probabilities. Stationarity is a reasonable condition on the correlations; it also serves as a first approximation in the nonstationary case. The next section describes some alternative approaches to modelling data of this type.

2. Alternative Approaches

One method for modelling time series of discrete-valued data is to approach the subject in a manner analogous to Box-Jenkins (1970); one such approach was considered by Jacobs and Lewis in two articles (1978a, 1978b). Jacobs and Lewis replace the linear combinations of the continuous-valued case with probabilistic mixtures and call their models discrete autoregressive-moving average (DARMA) models. As an example, consider the simplest Box-Jenkins model, the first-order autoregressive model (AR[1]), where the X_n process is formed according to:

$$X_n = \rho X_{n-1} + e_n ,$$

where $|\rho|<1$ and $\{e_n\}_{n=-\infty}^{\infty}$ are i.i.d. $(0,\sigma^2)$. The DAR(1) sequence, $\{D_n\}_{n=-\infty}^{\infty}$, is formed according to:

$$D_n = \begin{cases} D_{n-1} & \text{with probability } 0 \leq \rho \leq 1 \\ E_n & \text{with probability } 1 - \rho, \end{cases}$$

where $\{E_n\}_{n=-\infty}^{\infty}$ is an i.i.d. sequence of random variables in the discrete set. An analogue of the Yule-Walker equations exists in which linear combinations of correlations are replaced by probabilistic mixtures. An advantage in approaching the discrete case in a mode analogous to Box-Jenkins is that the directions which have and have not proven fruitful in the continuous-valued case may be of guidance. There are some disadvantages to the Jacobs-Lewis approach. For example, in both the DAR(1) and AR(1) models, ρ^j is the autocorrelation of lag j; however, ρ is forced to be nonnegative in the discrete case, so that all autocorrelation must be nonnegative, thereby restricting the model's applicability. In the DAR(1) model, D_{n-1} contains all information available at time n-1 about the past, as does X_{n-1} in the AR(1) model. However, in the discrete model there is randomizing between D_{n-1} and E_n; E_n contains no information about the past, and if E_n is chosen, then all memory before time n is lost forever. In other words, the memory of such discrete models is discontinuous; in the AR(1) model the memory dies out geometrically. These same advantages and disadvantages carry over to the general DARMA models.

Another alternative is the Markov chain or modifications thereof; no exposition of Markov chain modelling will be given here. Markov models of lagged dependency of two or more become quite complicated and cumbersome in terms of calculations. However, the Markov property, by its very construction, lends itself quite readily to forecasting.

Lomnicki and Zaremba (1955) and Kedem (1980b) model binary data as the clipping of an underlying process; a clipping is a truncation at zero. Kedem's work is for the situation where the underlying data is available and, if need be, can be properly centered; the original data, for computability reasons, is replaced by the coarse data, zeroes and ones, which are used for estimation. Lomnicki and Zaremba consider the clipping of the k^{th} difference of a Gaussian process.

3. Basic Framework

An alternative to defining the structure on the discrete process is to assume, as Lomnicki and Zaremba (1955) and Kedem (1980b) do, that it inherits a certain structure from an underlying continuous process. The discretization by a threshold or truncation of a continuous-valued process is a common phenomenon in engineering and biology. I consider a general procedure which includes thresholds and truncations as special cases.

A binary process, $\{D_n\}_{n=-\infty}^{\infty}$, is assumed to be generated by $\{X_n\}_{n=-\infty}^{\infty}$, a continuous, strictly stationary time series, and a monotone function $F : R \to [0,1]$ in the following way. Given $\{X_n\}_{n=-\infty}^{\infty}$, the D_n are independent and

$$\begin{aligned} P(D_n = 1|X_n) &= F(X_n) \\ P(D_n = 0|X_n) &= 1 - F(X_n). \end{aligned} \tag{3.1}$$

By the definition of D_n, let $0 < j_1 < j_2 < \ldots < j_s$ be integers, then

$$\begin{aligned} &P(D_1, D_{1+j_s}, D_{1+j_2}, \ldots D_{1+j_s} | X_1, X_{1+j_1}, X_{1+j_2}, \ldots, X_{1+j_s}) \\ &\quad = P(D_1|X_1) \cdot P(D_{1+j_1}|X_{1+j_1}) x \ldots x P(D_{1+j_s}|X_{1+j_s}). \end{aligned}$$

Dependence among the responses D_n, given X_n, could also be considered, although for most applications independence has intuitive appeal. The D_n process was generated through a response function F which maps R into [0,1]. Some examples of F follow.

Example 1. Truncation or Threshold Function

A continuous process is truncated at some value μ, so that, given X_n,

$$D_n = \begin{cases} 1 \text{ if } X_n \geq \mu \\ 0 \text{ if } X_n < \mu \end{cases}.$$

Example 2. Probability Cumulative Distribution Function

In this case, the greater the underlying value, the greater the probability of a one being generated. One can think of this as a random threshold or, equivalently, as the introduction of measurement error followed by truncation.

Example 3. Survival Functions

A survival function is one minus a probability c.d.f.; this has the reverse effect of Example 2.

Example 4. Defective c.d.f.

This puts bounds on the response probabilities below one and/or above zero.

The three examples given in the Introduction can be discussed in terms of these generating procedures. In the mood analysis and business cycle turning points examples, the underlying processes are our true mood and the actual value of the economic indicator, respectively. In the labor force participation example, the underlying process could be the difference in the lifetime utility at time n of the individual if employed and if not employed (Heckman (1979)).

A binary time series generated by these methods is strictly stationary, since the underlying X_n process is strictly stationary. The probability of the event $\{D_1 = d_1, D_2 = d_2, \ldots, D_n = d_n\}$, where $(d_1, d_2, \ldots, d_n) \in \{0,1\}^n$, is

$$P\{D_1 = d_1, D_2 = d_2, \ldots, D_n = d_n\} =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}[F(x_1)]^{d_1}[1 - F(x_1)]^{1-d_1}\cdots[F(x_n)]^{d_n}[1-F(x_n)]^{1-d_n} d\, G_n(x_1,x_2,\ldots,x_n). \quad (3.2)$$

Second-order stationarity of the underlying process will not insure second-order stationarity of the discrete process. A characterization lemma, which follows, shows that all strictly stationary binary processes can be generated by the above procedure through just the Gaussian processes. The integral in

equation (3.2) is similar to a convolution and can be given probabilistic meaning. The following lemma, a variation of a known result (Feller (1971), p. 144), gives an alternative view of the response function F. Define the sets $C(0) = (-\infty,0)$ and $C(1) = [0,\infty)$.

Lemma 3.1

Let the X_s process and F be as above and let G_n be the distribution function of $(X_1,X_2,\ldots,X_n)$. Let $\{Y_n\}_{i=-\infty}^{\infty}$ be i.i.d. with c.d.f. F, independent of $\{X_n\}_{i=-\infty}^{\infty}$, and define $V_i = X_i - Y_i$. Then, for $(d_1,d_2,\ldots,d_n) \in \{0,1\}^n$,

$$P\{D_1 = d_1,\ D_2 = d_2,\ldots,D_n = d_n\} = P\{V_1 \in C(d_1),\ V_2 \in C(d_2),\ldots,V_n \in C(d_n)\}$$

Proof. The event $\{V_1 \in C(d_1),\ V_2 \in C(d_2),\ldots,V_n \in C(d_n)\}$

$$= \{Y_j \leq X_j \text{ for } j \ni d_j = 1 \text{ or } X_j < Y_j \text{ for } j \ni d_j = 0;\ j = 1,2\ldots,n\}.$$

Pick $\varepsilon > 0$; $R = \bigcup_{k=-\infty}^{\infty} I_k$ where $I_k = (k\varepsilon,(k+1)\varepsilon]$. Since the $\{Y_i\}$ are independent and independent of the X_i process, the probability of the above event is approximately

$$\sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \cdots \sum_{k_n=-\infty}^{\infty} \left\{ \prod_{j=1}^{n} [F(x_j)]^{d_j} [1 - F(x_j)]^{1-d_j} \right\} \cdot P\{X_1 \in I_{k_1},\ldots,X_n \in I_{k_n}\},$$

where $x_j \in I_{k_j}$.

If $x \in I_k = (k\varepsilon,(k+1)\varepsilon]$, then $F(k\varepsilon) \leq F(x) \leq F((k+1)\varepsilon)$, and the above probability is bounded above and below by sums which depend only on ε and which both converge to $P\{D_1 = d_1,\ D_2 = d_2,\ldots,D_n = d_n\}$, defined by expression (3.2), as $\varepsilon \to 0$. Q.E.D.

This lemma shows that there is a duality with which one can view the generation of discrete processes through response functions. One can view the binary process as being generated by the X_n process and F or, equivalently, by the V_n process and a truncation at zero. Note that if X_n is an ARMA(p,q) process,

then V_n will usually not be an ARMA process of the same order, although it will be in the ARMA family (see Granger and Morris (1976)). Lemma 3.1 says that probabilitistically we need only consider truncation as the discretization mechanism if we are only concerned with properties of strictly stationary time series. The binary process is a strictly stationary time series with

$$E(D_n) = P(V_n \geq 0) = P(D_n = 1)$$

$$Var(D_n) = P(V_n \geq 0) \cdot P(V_n < 0)$$

and

$$Cov(D_n, D_{n+j}) = P(V_n \geq 0, V_{n+j} \geq 0) - [P(V_n \geq 0)]^2, \quad j \geq 1.$$

If $\{X_n\}_{n=-\infty}^{\infty}$ is a Gaussian process with $E(X_n) = 0$, $Var(X_n) = \tau^2$, correlation structure $\{\rho(j)\}_{j=1}^{\infty}$, and if $F = \phi_{0,b^2}$, the c.d.f. of $N(0,b^2)$, then the V_n process is Gaussian with $E(V_n) = 0$, and the covariance matrix of $(V_{n+1}, V_n, \ldots, V_1)$ is given by M_n with elements

$$\sigma_{ij} = \begin{cases} \rho(|i-j|)\tau^2 & \text{if } i \neq j \\ (1+k)\tau^2 & i = j \end{cases} \quad i, j = 1,2,\ldots,n+1,$$

where $k = b^2/\tau^2$ is the ratio of the variance of the response function to that of the underlying process. If there is no response function variation, i.e., a truncation, then $k = 0$; as the amount of response function variation increases, the autocorrelations of the V_n process are reduced, so that the past of the V_n process has less effect on its future than the past of the X_n process has on the future of the X_n process. Defining $\underline{\rho}_n' = (\rho(1),\rho(2),\ldots,\rho(n))$ and $\underline{V}_n = (V_n, V_{n-1}, \ldots, V_1)$ we have

$$M_n = \tau^2 \left[\begin{array}{c|c} (1+k) & \underline{\rho}_n' \\ \hline \underline{\rho}_n & M_{n-1} \end{array}\right].$$

The regression coefficients of V_{n+1} on $(V_n, V_{n-1}, \ldots, V_1)$ are

$$\underline{S}_n' = \underline{\rho}_n' \, M_{n-1}^{-1} \tag{3.3}$$

and, defining $Z_n = \underline{S}_n' \, \underline{V}_n$, Z_n is Normal with mean zero and variance

$$Q_n = \tau^2 \, \underline{\rho}_n' \, M_{n-1}^{-1} \, \underline{\rho}_n \, . \tag{3.4}$$

Throughout, τ^2 will be assumed to be one; the effect of taking other values for τ^2 will be pointed out.

The bivariate and trivariate probabilities involving the D_n process can be calculated, since the process can be viewed as a truncation of the V_n process. Classical results on the probabilities of bivariate and trivariate normal quadrants (orthants) can be used; see, for instance, the survey article by S.S. Gupta (1963). Applying these results to $\underline{V}_n$, we have for $j, \ell \geq 1$, $(d_n, d_{n+j}) \in \{0,1\}^2$, $(d_n, d_{n+j}, d_{n+j+\ell}) \in \{0,1\}^3$

$$P\{D_n = d_n, \, D_{n+j} = d_{n+j}\} = 1/4 + (-1)^{d_n + d_{n+j}} \frac{\arcsin\left[\frac{\rho(j)}{1+k}\right]}{2\Pi} \tag{3.5}$$

and

$$\begin{aligned} P\{D_n = d_n, \, D_{n+j} = d_{n+j}, \, D_{n+j+\ell} = d_{n+j+\ell}\} \\ = 1/8 + (-1)^{d_n + d_{n+j}} \frac{\arcsin\left[\frac{\rho(j)}{1+k}\right]}{4\Pi} + (-1)^{d_n + d_{n+j+\ell}} \frac{\arcsin\left[\frac{\rho(j+\ell)}{1+k}\right]}{4\Pi} \\ + (-1)^{d_{n+j} + d_{n+j+\ell}} \frac{\arcsin\left[\frac{\rho(\ell)}{1+k}\right]}{4\Pi} \, . \end{aligned} \tag{3.6}$$

The correlation structure of the D_n process is

$$\delta(j) = \frac{2\arcsin\left[\frac{\rho(j)}{1+k}\right]}{\Pi} \qquad j \geq 1. \tag{3.7}$$

The discretization operates on the underlying correlations by rescaling in two stages, first by scaling down by the factor (1 + k), due to the response function, then by applying the scaled arcsin function, due to the truncation. Note that k does not depend on $\{\rho(j)\}_{j=1}^{\infty}$ except that any such structure not compatible with a finite τ^2 constrains k to be zero. The function $2\arcsin[\cdot]/\pi$ is a monotone increasing function of [-1, 1] onto itself, shrinking the value towards zero.

The above derivations of the bivariate and trivariate probabilities of the D_n process were possible because of closed-form expressions for 2 and 3 dimensional multivariate normal orthant probabilities. For dimensions greater than 3 there is no general expression. There are a few special cases in 4 dimensions; see Gupta (1963), and Cheng (1969). Numerical methods are described in Abrahamson (1964), McFadden (1956), and Moran (1956), and some recursive formulas in Schlafi (1858), David (1953), Sondhi (1961), and Choi (1975). Only in the equicorrelated case is a closed-form expression known for general n (Gupta, 1963). If good approximations to the n-dimensional probabilities were available, then the joint and one step-ahead transition probabilities of the D_n process could be approximated.

The following lemma is a characterization of those binary processes which can be generated by an underlying process through a response function. All strictly stationary binary processes can be generated in this manner; in fact, all can be generated under the Gaussian assumptions.

Lemma 3.2

Let $\{D_n\}_{n=-\infty}^{\infty}$ be a binary-valued strictly stationary time series with finite dimensional joint probabilities $P\{D_1 = 0, D_2 = 0, \ldots, D_n = 0\}$ specified for $n \geq 1$, then there exist an underlying stationary Gaussian process $\{X_n\}_{n=-\infty}^{\infty}$ and a

Gaussian c.d.f. $F:R\to[0,1]$ such that the joint probabilities are given by expression (3.2).

Proof

We can assume without loss of generality that $Var(X_n) = 1$, $E(X_n) = 0$, and the mean and variance of F are μ and k, respectively. Define $C = \mu/\sqrt{1 + k}$, so that

$$P\{D_1 = 0\} = P\{V_1 \leq 0\} = \Phi_{0,1}(c)$$

and $c = \Phi_{0,1}^{-1}(P\{D_1 = 0\})$ is uniquely determined. Suppose that $\rho(1),\rho(2),\ldots,\rho(j-1)$ have been determined, and let Σ_j be the correlation matrix with elements $\rho_{\ell,m} = \rho(|\ell-m|)$, $\ell,m = 1,2,\ldots,j+1$, with Σ_j, therefore, being only a function of $\rho(j)$. Let $\underline{0}'_n = (0,0,\ldots,0)$ and $\underline{1}'_n = (1,1,\ldots,1)$ be n-dimensional vectors of zeroes and ones. Then we have the following:

$$\lim_{\rho(j)\uparrow+1} \Phi_{\underline{0}_{j-1},\Sigma_j}(c\,\underline{1}_{j-1}) = P\{V_1 \leq c,\ V_2 \leq c,\ldots,V_j \leq c,\ V_1 \leq c\}$$
$$= P\{D_1 = 0,\ldots,D_j = 0\}$$
$$\lim_{\rho(j)\downarrow-1} \Phi_{\underline{0}_{j-1},\Sigma_j}(c\,\underline{1}_{j-1}) = P\{V_1 \leq c,\ V_2 \leq c,\ldots,V_j \leq c,\ V_1 > c\}$$
$$= 0$$

The distribution $\Phi_{\underline{0}_{j-1},\Sigma_j}(c\,\underline{1}_{j-1})$ is continuous in $\rho(j)$, and Slepian's Theorem (see Slepian (1962), Das Gupta et al. (1972)) says that $\Phi_{\underline{0}_{j-1},\Sigma_j}(c\underline{1}_{j-1})^{II}$ is a monotone increasing function of $\rho(j)$. Therefore, there exists a $\rho(j)$ such that $-1 \leq \rho(j) \leq 1$ and

$$\Phi_{\underline{0}_{j-1},\Sigma_j}(c\,\underline{1}_{j-1}) = P\{D_1 = 0,D_2 = 0,\ldots,D_j = 0,D_{j+1} = 0\},$$

and by mathematical induction there exists a sequence $\{\rho(j)\}_{j=1}^{\infty}$. By Kolmogorov's Existence Theorem there exists a stationary Gaussian process with correlation structure $\{\rho(j)\}_{j=1}^{\infty}$. Q.E.D.

A result more general than Slepian's Theorem is available for the family of elliptical distributions (See Das Gupta et al. (1972)).

The next section is concerned with predicting and/or determining the conditional distribution of D_{n+1}, given data $D_1,D_2,\ldots,D_n$.

4. The Binary Prediction Model

For a time series of discrete data an important problem is how to use the data in making decisions about the future behavior of the series. One may want to predict this future behavior or make probabilistic statements about it. To be more precise, given $d_1,d_2,\ldots,d_n$, these problems consist of predicting and/or determining the conditional distribution of d_{n+j}, $j \geq 1$. In terms of the three examples given in the Introduction, the data $(d_1,d_2,\ldots,d_n)$, would consist of employment history, moods, and economic directional changes, and our objective would be to predict and/or determine the distribution of future employment, mood, and directional change. The determination of the transitional probabilities $P(D_{n+1} = d_{n+1}|D_n = d_n,\ldots,D_1 = d_1)$ are of fundamental importance. The derivation of approximations to these conditional probabilities is discussed in Keenan (1980) in the context of the loss of Markov property; the difference between the actual probability and certain reasonable approximations can be viewed as a measure of such a loss. Therefore, the determination of the conditional distribution of D_{n+1}, given $D_1,\ldots,D_n$, will not be discussed here, but rather the related problem of predicting D_{n+1}, given $D_1,D_2,\ldots,D_n$.

It is assumed that $\{X_n\}_{n=-\infty}^{\infty}$ is a strictly stationary time series generating $\{D_n\}_{n=-\infty}^{\infty}$ through a response function $F:R\to[0,1]$, as follows:

Given $\{X_n\}_{n=-\infty}^{\infty}$,

$$P(D_n = 1|X_n) = F(X_n)$$

$$P(D_n = 0|X_n) = 1 - F(X_n) \quad .$$

Given data $d_1, d_2,\ldots,d_n$ from this process, the problem considered in this section is to predict d_{n+j}. For simplicity we shall deal with $j = 1$. To predict d_{n+1}, a prediction accuracy measure is needed. Since in binary data there is only one possible wrong and right prediction, the probability of predicting incorrectly is a reasonable measure of prediction uncertainty. The probability of error using $\tilde{D}_{n+1}$ is given by

$$P(\tilde{D}_{n+1} \neq D_{n+1}) = \int_{\{\tilde{D}_{n+1} \neq D_{n+1}\}} dP_{(D_1,D_2,\ldots,D_n,D_{n+1})}(d_1,d_2,\ldots,d_n,d_{n+1}), \tag{4.1}$$

where $\tilde{D}_{n+1}$ is a predictor of D_{n+1} based on the data $d_1,d_2,\ldots,d_n$. The broadest class of predictors which can be considered is the class of all randomized rules. For an arbitrary, randomized rule δ, the probability of error, P.E.(δ), is:

$$\begin{aligned} P.E.(\delta) := \sum_{(d_1,d_2,\ldots,d_n)} & [\delta(d_1,d_2,\ldots,d_n)\, P\{D_1 = d_1, D_2 = d_2,\ldots,D_n = d_n, D_{n+1} = 0\} \\ & + (1 - \delta(d_1,d_2,\ldots,d_n))P\{D_1 = d_1, D_2 = d_2,\ldots,D_n = d_n, D_{n+1} = 1\}]. \end{aligned} \tag{4.2}$$

One nonrandomized rule which is a reasonable predictor of D_{n+1} is

$$\hat{D}_{n+1} = \begin{cases} 1 \text{ if } P(D_{n+1} = 1 | D_n = d_n,\ldots,D_1 = d_1) \geq 1/2 \\ 0 \text{ if } P(D_{n+1} = 1 | D_n = d_n,\ldots,D_1 = d_1) < 1/2, \end{cases} \tag{4.3}$$

which predicts that value which is the highest probable. A randomized rule which is just as reasonable is that which randomizes with the conditional probability:

$$\tilde{\delta}(d_1,d_2,\ldots,d_n) = P(D_{n+1} = 1 | D_n = d_n,\ldots,D_1 = d_1).$$

The next lemma shows that $\hat{D}_{n+1}$, a nonrandomized rule, has minimum probability of error among all randomized rules.

Lemma 4.1

$\hat{D}_{n+1}$ has minimum probability of error within the class of all randomized rules.

Proof - Define an arbitrary randomized rule δ by $\delta(d_1,d_2,\ldots,d_n) = P(1|d_1,d_2,\ldots,d_n)$. The P.E.($\delta$) is given by equation (4.2). For a given $(d_1,d_2,\ldots,d_n)$,

$$\delta(d_1,d_2,\ldots,d_n)\ P\{D_1 = d_1,\ldots,D_n = d_n,\ D_{n+1} = 0\} +$$
$$(1 - \delta(d_1,\ldots,d_n))\ P\{D_1 = d_1,\ldots,D_n = d_n,\ D_{n+1} = 1\}$$

lies between $P\{D_1 = d_1,\ldots,D_n = d_n,\ D_{n+1} = 0\}$ and $P\{D_1 = d_1,\ldots,D_n d_n,\ = D_{n+1} = 1\}$, and so the P.E. is minimized by

$$\delta^*(d_1,d_2,\ldots,d_n) = \begin{matrix} 0 \text{ if } P(D_1 = d_1,\ldots,D_n = d_n,\ D_{n+1} = 1) < P(D_1 = d_1,\ldots,D_n = d_n, D_{n+1} = 0) \\ 1 \text{ if } P(D_1 = d_1,\ldots,D_n = d_n,\ D_{n+1} = 1) < P(D_1 = d_1,\ldots,D_n = d_n, D_{n+1} = 0), \end{matrix}$$

which is equivalent to $\hat{D}_{n+1}$. Q.E.D.

This lemma shows that if our main concern is predicting D_{n+1}, then we need not know the exact probability, $P(D_{n+1} = 1|d_1,d_2,\ldots,d_n)$, but only whether or not it exceeds or equals 1/2.

We saw in Section 3 that, when $\{X_n\}_{n=-\infty}^{\infty}$ and $F = \Phi_{(0,b^2)}$ are Gaussian, the D_n are binomial (1, 1/2), with the bivariate and trivariate probabilities given by expressions (3.5) and (3.6), respectively. Consider the special case of just one data point, d_1. The minimum probability of error predictor of D_2 is given by

$$\hat{D}_2 = \begin{cases} 1 \text{ is } P(D_2 = 1|D_1 = d_1) \geq 1/2 \\ 0 \text{ is } P(D_2 = 1|D_1 = d_1) < 1/2, \end{cases}$$

where

$$P(D_2 = d_2|D_1 = d_1) = 1/2 + (-1)^{d_1 + d_2} \frac{\arcsin\left[\frac{\rho(1)}{1+k}\right]}{\pi}.$$

The predictor of D_2 is what we intuitively expect, namely

$$\hat{D}_2 = \begin{cases} d_1 & \text{if } \rho(1) > 0 \\ 1-d_1 & \text{if } \rho(1) < 0. \end{cases}$$

The P.E.$(\hat{D}_2)$ can be calculated:

$$P(\hat{D}_2 \neq D_2) = 1/2 - \frac{\arcsin\,[|\rho(1)|/(1+k)]}{\Pi} \quad . \tag{4.4}$$

For $\rho(1) = 0$, we have independence of D_2 and D_1, and since D_1 gives no information about D_2, the probability of predicting wrongly is 1/2. If $|\rho(1)| = 1$, then $\mathrm{Var}(X_n) = \tau^2$ is infinite, $k = b^2/\tau^2 = 0$, and the probability of error is zero, as expected.

Consider the case of two data points, d_1, d_2. Before considering $\hat{D}_3$, look at the following two predictors, $\hat{D}_3^{(j)}$, $j = 1,2$, given by

$$\hat{D}_3^{(j)} = \begin{cases} 1 \text{ if } P(D_3 = 1|D_{3-j}) \geq 1/2 \\ 0 \text{ if } P(D_3 = 1|D_{3-j}) < 1/2 \end{cases} \quad . \tag{4.5}$$

$\hat{D}_3^{(1)}$ and $\hat{D}_3^{(2)}$ are predictors based on only part of the available data, one and two time points back, respectively. From the case of prediction based on one data point, we know that for $j = 1, 2$

$$\text{P.E.}(\hat{D}_3^{(j)}) = 1/2 - \frac{\arcsin\,[|\rho(j)|/(1+k)]}{\Pi} \tag{4.6}$$

Now consider $\hat{D}_3$ which uses both data points. Using equations (3.5) and (3.6), the conditional probabilities of D_3 given D_1 and D_2 are explicitly determined. Figure 1 divides the $(\rho(1), \rho(2))$ plans into 4 parts via the equiangular lines.

Figure 1

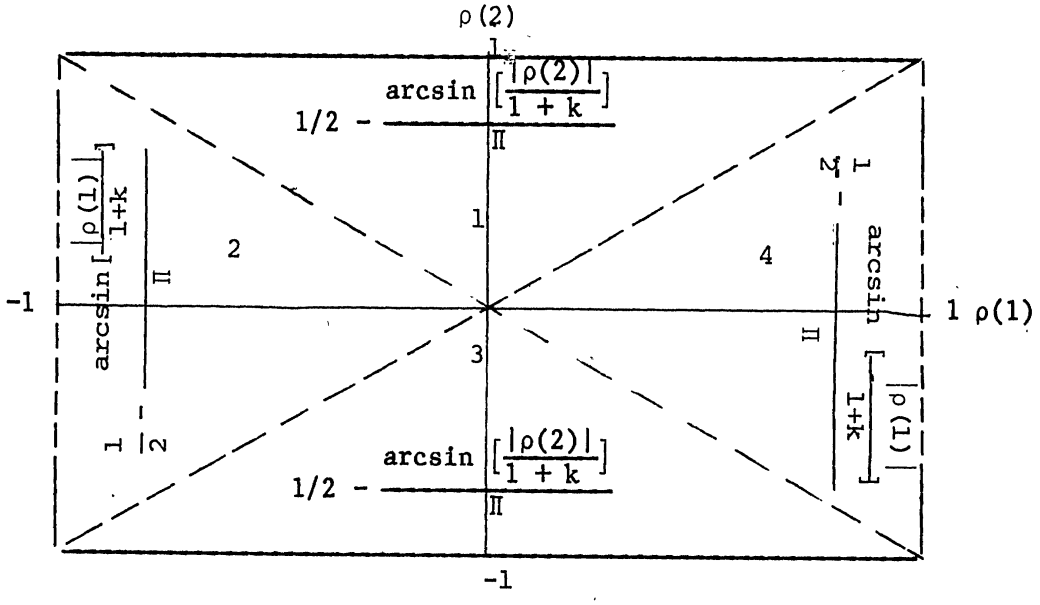

Table 1 shows $\hat{D}_3$ for these four areas, and Figure 1 also plots the probability of error of $\hat{D}_3$ as a function of $(\rho(1), \rho(2))$.

Table 1

Data	Area			
	1	2	3	4
$D_1=1,\ D_2=1$	$\hat{D}_3=1$	$\hat{D}_3=0$	$\hat{D}_3=0$	$\hat{D}_3=1$
$D_1=0,\ D_2=1$	$\hat{D}_3=0$	$\hat{D}_3=0$	$\hat{D}_3=1$	$\hat{D}_3=1$
$D_1=1,\ D_2=0$	$\hat{D}_3=1$	$\hat{D}_3=1$	$\hat{D}_3=0$	$\hat{D}_3=0$
$D_1=0,\ D_2=0$	$\hat{D}_3=0$	$\hat{D}_3=1$	$\hat{D}_3=1$	$\hat{D}_3=0$

Figure 1 suggests that if $|\rho(1)|>|\rho(2)|$, use only the preceding observation to predict; if $|\rho(2)|>|\rho(1)|$, use the penultimate observation. In each case, if the larger correlation is positive, predict concordance with the past value; if negative, predict discordance. This is summarized in the following lemma.

Lemma 4.2

For n = 2 data points when the X_s process and $F = \Phi_{(0,b^2)}$ are both Gaussian, the minimum probability of error predicator of D_3 is $\hat{D}_3^{(\ell)}$, where $|\rho(\ell)|=\text{Max}\ \{|\rho(1)|, |\rho(2)|\}$ and $\hat{D}_3^{(\ell)}$ is given by expression (4.5). The minimum probability of error is

$$1/2 - \frac{\arcsin\ [|\rho(\ell)|/1+k]}{\Pi} .$$

Remark. This lemma states that for two data points, the minimum probability of error prediction is based on only one of the two points, that which has highest correlation with D_3; knowledge of the other point does not help.

Proof. From Table 3 it can be seen that $\hat{D}_3$ equals $\hat{D}_3^{(1)}$ in areas 2 and 4 and $\hat{D}_3^{(2)}$ in areas 1 and 3. Therefore, in areas 2 and 4 the P.E.$(\hat{D}_3)$ is

$$1/2 - \frac{\arcsin[|\rho(1)|(1+k)]}{\Pi} ,$$

and in areas 1 and 3 it is

$$1/2 - \frac{\arcsin[|\rho(2)|/(1+k)]}{\Pi} . \qquad \text{Q.E.D.}$$

For an arbitrary $n \geq 2$, define $\hat{D}_{n+1}^{(1,2)}$ as

$$\hat{D}_{n+1}^{(1,2)} = \begin{cases} 1 \text{ if } P(D_{n+1} = 1 | D_n = d_n, D_{n-1} = d_{n-1}) \geq 1/2 \\ 0 \text{ if } P(D_{n+1} = 1 | D_n = d_n, D_{n-1} = d_{n-1}) < 1/2, \end{cases}$$

(i.e., optimal predictor based on only the two preceding values), and let $\hat{\hat{D}}_{n+1}^{(1,2)}$ be any decision rule based on only the two preceding values.

Corrollary 4.3.

If the X_s process and $F = \Phi_{(0,b^2)}$ are Gaussian, then for an arbitrary $n \geq 2$,

$$\begin{aligned} \text{P.E.}(\hat{\hat{D}}_{n+1}^{(1,2)}) &\geq \text{P.E.}(\hat{D}_{n+1}^{(1,2)}) \\ &= 1/2 - \frac{\arcsin[\text{Max}\{|\rho(1)|, |\rho(2)|\}/(1+k)]}{\Pi} \\ &\geq \text{P.E.}(\hat{D}_{n+1}) . \end{aligned}$$

Proof. The P.E.($\hat{\hat{D}}^{(1,2)}_{n+1}$) is given by expression (4.2). Since $\hat{\hat{D}}^{(1,2)}_{n+1}$ does not depend on $(d_1,d_2,\ldots,d_{n-1})$ and the D_n process is strictly stationary, the P.E.($\hat{\hat{D}}^{(1,2)}_{n+1}$) reduces to

$$\sum_{(d_{n-1},d_n,d_{n+1})} P\{D_{n-1} = d_1,\ D_n = d_n,\ D_{n+1} = d_{n+1},\ \hat{\hat{D}}^{(1,2)}_{n+1} = 1 - d_{n+1}\}$$

which is greater than or equal to P($\hat{D}^{(1,2)}_{n+1}$), which by Lemma 4.2 [and since $\hat{D}^{(1,2)}_{n+1}$ does not depend on $(d_1,\ldots,d_{n-1})$], is

$$1/2 - \frac{\arcsin[\text{Max}\{|\rho(1)|,|\rho(2)|\}\ (1+k)]}{\Pi},$$

which is greater than or equal to P.E.($\hat{D}_{n+1}$), by definition of $\hat{D}_{n+1}$. Q.E.D.

Remark. Therefore, for data, $d_1,d_2,\ldots,d_n$, $n \geq 2$, we have a lower bound on how well predictors can do using only the present and penultimate observations and an upper bound on the optimal predictors which use more than the previous two preceding observations.

For the remainder of this section we will consider the binary prediction problem for an arbitrary $n \geq 2$ and the most fundamental model, that of the first-order autoregressive. Therefore, the X_s process is Gaussian AR(1), $X_s = \rho X_{s-1} + e_s$, $|\rho| < 1$, $\text{Var}(X_s) = 1$ and $F = \Phi_{(0,k)}$. We have seen that the D_s process generated by X_s and F is not a Markov chain. An important question is: How much of the past must be used for prediction purposes? Previously we considered the binary prediction problem for the arbitrary Gaussian case with n = 2 data points. The determination of the conditional distribution of D_3, given D_2 and D_1, allowed for explicit probability of error calculations. For $n \geq 3$, general closed-form expressions for the conditional distribution of

D_{n+1} given $D_n, D_{n-1}, \ldots, D_1$ are not available. In Keenan (1980) six approximations, A(1) to A(6), to this distribution are developed in a different context, that of determining the loss of Markov property due to discretization. Rather than attempting to numerically approximate n and (n + 1) dimensional integrals, the approach of the six approximations is to reduce the dimensions to 2 and 3 for which closed form expressions, (3.5) and (3.6), are available. Approximations A(2) and A(3) are the conditional distributions of D_{n+1}, given D_n and D_{n-1} and given D_n, respectively. Approximations A(4), A(5), and A(6) use the binary data $(D_n, D_{n-1}, \ldots, D_1)$ first, to estimate the underlying unobserved data $(V_n, V_{n-1}, \ldots, V_1) = \underline{V}'_n$; then to estimate $Z_n = \underline{S}'\underline{V}_n$, where $\underline{S}'_n$ is given by expression (3.3); and finally to calculate the conditional distribution of D_{n+1} given the estimate of Z_n. In Keenan (1980), or using Kalman filtering methods (see Jazwinski (1970), Duncan and Horn (1972) Downing et al. (1980) etc.), recursive expressions for $\underline{S}'_n = (S_{n,1}, S_{n,2}, \ldots, S_{n,n})$, $Q_n = \mathrm{Var}(Z_n)$, and $R_n = k/(1+k-Q_n)$ are

$$S_{n,j} = \begin{cases} \rho R_{n-1}\, S_{n-1,j-1} & 2 \leq j \leq n \\ \rho(1-R_{n-1}) & j = 1 \end{cases} \qquad (4.7)$$

$$Q_n = \rho^2(1 - R_{n-1}) + \rho^2 R_{n-1}\, Q_{n-1}.$$

The factor $(1-R_{n-1})$ is the usual Kalman gain. Approximations A(4), A(5), and A(6) estimate Z_n by

$$\begin{aligned} Z_n^{(4)} &= \sum_{j=1}^{n} S_{n,j} \times E(V_j | D_j = d_j) = \Big[\sum_{j=1}^{n} (-1)^{1-d_{n-j+1}} S_{n,j}\Big]\sqrt{\frac{2(1+k)}{\Pi}} \\ Z_n^{(5)} &= S_{n,1} \times E(V_n | D_n = d_n,\ D_{n-1} = d_{n-1}) \\ &\quad + S_{n,2} \times E(V_{n-1} | D_n = d_n,\ D_{n-1} = d_{n-1}) \end{aligned} \qquad (4.8)$$

$$= [(-1)^{1-d_n} S_{n,1} + (-1)^{1-d_{n-1}} S_{n,2})] \left(\sqrt{\frac{2(1+k)}{\Pi}}\right) \times \frac{1 + (-1)^{d_n+d_{n-1}}[\rho/(1+k)]}{1 + (-1)^{d_n+d_{n-1}} \frac{\arcsin[\rho/(1+k)]}{\Pi}} \quad (4.9)$$

and

$$Z_n^{(6)} = S_{n,1} E(V_n|D_n)$$
$$= \left((-1)^{1-d_n} S_{n,1}\right) \sqrt{\frac{2(1+k)}{\Pi}} . \quad (4.10)$$

Approximation A(1) is the conditional distribution of D_{n+1} given D_n and E_{n-1}; i.e.,

$$P(D_{n+1} = 1|D_n = d_n, E_{n-1} \varepsilon e_{n-1}) ,$$

where E_{n-1} is the binary (0,1) function of $(D_{n-1},\ldots,D_1)$ defined by

$$E_{n-1} = \begin{cases} 1 \text{ if } Z_{n-1} \geq 0 \\ 0 \text{ if } Z_{n-1} < 0 \end{cases}$$

and

$$e_{n-1} = \begin{cases} \{0,1\} & \text{if Range } (Z_{n-1}) = R \\ 1 & \text{if Range } (Z_{n-1}) = [0,\infty] \\ 0 & \text{if Range } (Z_{n-1}) = [-\infty,0] \end{cases} .$$

Justification for this approximation is described in Keenan (1980). The correlations of (D_{n+1}, D_n, E_{n-1}) are

$$\text{Corr}(D_{n+1}, D_n) = \frac{2\arcsin[\rho/(1+k)]}{\Pi}$$

$$\text{Corr}(D_n, E_{n-1}) = \frac{2\arcsin[\text{sign}(\rho)\cdot(Q_{n-1}/(1+k))^{1/2}]}{\Pi}$$

and

$$\text{Corr}(D_{n+1}, E_{n-1}) = \frac{2\arcsin[|\rho|\cdot(Q_{n-1}/(1+k))^{1/2}]}{\Pi} .$$

Approximations A(1) and A(4) were shown to be better than the others; these were the only approximations which used all of the data ($D_1 = d_1$, $D_2 = d_2,\ldots,D_n = d_n$).

Consider the predictors, $\hat{D}_{n+1}^{A(j)}$, $j = 1,2,\ldots,6$, where the approximations A(1) to A(6) are substituted for $P(D_{n+1} = 1 | D_n = d_n,\ldots,D_1 = d_1)$ in expression (4.3). The next theorem allows us to calculate the probability of error for these predictors.

Theorem 4.4

If the D_s process is generated by an X_s process which is Gaussian AR(1), $|\rho| < 1$, $\mathrm{Var}(X_s) = \tau^2$, and response function $F = \Phi_{(0,b^2)}$ with $k = b^2/\tau^2$, then for n data points

$$P.E.(\hat{D}_{n+1}^{A(2)}) = P.E.(\hat{D}_{n+1}^{A(3)}) = P.E.(\hat{D}_{n+1}^{A(5)}) = P.E.(\hat{D}_{n+1}^{A(6)})$$

$$= 1/2 - \frac{\arcsin[|\rho|/(1+k)]}{\pi}$$

$$= \begin{cases} P.E.(\hat{D}_{n+1}^{A(1)}) \text{ if } Q_{n-1} \leq 1/(1+k) & (4.11) \\ P.E.(\hat{D}_{n+1}^{A(4)}) \text{ if } \rho r \leq 1/2. & (4.12) \end{cases}$$

Proof. Without loss of generality we can assume that ρ is positive; the same conditions hold for ρ negative. By strict stationarity $P.E.(\hat{D}_{n+1}^{A(2)}) = P.E.(\hat{D}_3^{(1,2)})$ and $P.E.(\hat{D}_{n+1}^{A(3)}) = P.E.(\hat{D}_3^{(1)})$; Corollary 4.3 now gives the result for A(2) and A(3). Using expression (4.7) and comparing expressions (4.9) and (4.10) the result follows for A(5) and A(6). For approximations A(4), A(5), and A(6), $Z_n^{(j)} \geq 0$ is equivalent to $A(j) \geq 1/2$, $j = 4,5,6$. For A(4) the most extreme case is where D_n is one (zero) and the entire infinite past $(D_{n-1},D_{n-2},\ldots,D_1,D_0,D_{-1},\ldots)$ is all zeroes (ones); $Z_n^{(4)}$ is greater than or equal to zero in the case if $\rho r \leq 1/2$. If the data only goes back finitely and $\rho r \leq 1/2$, then $Z_n^{(4)} \geq 0$ is certainly satisfied. Approximation A(1) is the same as A(3) except when the past differs

from the present ($d_1 = d_2 = \ldots = d_{n-1} \neq d_n$); but A(4) and A(3) will both be greater than or equal to 1/2 at these extreme points if $Q_{n-1} \leq 1/(1+k)$. Q.E.D. Conditions $Q_{n-1} \leq 1/(1+k)$ and $\rho r \leq 1/2$ are conditions measuring the correlation of D_n with D_{n+1} relative to that of $(D_{n-1},D_{n-2},\ldots,D_1)$ with D_{n+1}, and $(D_{n-1},D_{n-2},\ldots,D_1,D_0,\ldots)$ with D_{n+1}, respectively. Tables 2 and 3 show values for ρ and k for which the conditions of Theorem 4.4 are satisified.

Table 2

Conditions for expressions (4.11) and (4.12) to be satisfied in terms of $|\rho|$ and K.

	$Q_m \leq 1/(1+k)$ (for all m)	p . r < 1/2
k	$\lvert\rho\rvert$	$\lvert\rho\rvert$
= 0	$\leq$ 1.00	$\leq$ 1.00
$\leq$ .2	$\leq$ .95	$\leq$ .95
$\leq$ .5	$\leq$ .90	$\leq$ .85
$\leq$ 1.0	$\leq$ .85	$\leq$ .80
$\leq$ 2.0	$\leq$ .80	$\leq$ .70

The minimum Probability of Error predictor involves $P(D_{n+1}=1|D_n=D_n,\ldots,D_1=d_1)$. Rather than numerically approximating the probabilities of the exact n+1 and n-dimensional random variables, the approximations A(1) to A(6) calculate exactly the probabilities of 3 and 2 dimensional approximations to the n+1 and n dimensional random variables. In Keenan (1980) different approximations are shown to do well for different parameter values; for any given parameter region, either A(1) or A(4) appears to be a reasonable approximation to the probability $P(D_{n+1} = 1|D_n=d_n,\ldots,D_1=d_1)$. Therefore, Corollary 4.3 and Theorem 4.4 suggest that if the underlying process is AR(1) (therefore, Markov) and D_n is a discretization of this process, then even though D_n is not a Markov chain (having in

fact infinite memory), for prediction purposes we should act as if D_n is a Markov chain and predict from the present value with

$$1/2 - \frac{\arcsin[|\rho|/(1+k)]}{\pi}$$

as the approximate probability of error for one-step-ahead prediction.

5. Multinomial Time Series

The previous sections have dealt with binary data; both the general framework and the specific results under Gaussian assumptions were for the binary case. We can now consider a more general model for the generation of ordered categories-multinomial time series by an underlying continuous valued-state space process and a response function F. Asymmetric Gaussian randomization, $F = \Phi_{(\mu,b^2)}$ and μ not equal to zero, is a special case. Each of the three examples of binary times series given in the Introduction can be generalized to the multinomial case; for example, one could consider more than 2 psychological states and more than the simple dichotomy of being employed or unemployed.

For $m \geq 1$, an $(m + 1)$ valued-multinomial process, $\{D_n\}_{n=-\infty}^{\infty}$, is assumed to be generated by an underlying continuous-valued, strictly stationary time series, $\{X_n\}_{n=-\infty}^{\infty}$, $E(X_n) = 0$, a probability c.d.f. $F : R \to [0,1]$, and the sequence $-\infty = \mu_{-1} \leq \mu_0 \leq \cdots \leq \mu_{m-1} \leq \mu_m = +\infty$, such that, given $\{X_n\}_{n=-\infty}^{\infty}$ the D_n are independent with

$$D_n = \{j \text{ with probability } F(X_n - \mu_{j-1}) - F(X_n - \mu_j), \qquad j=0,1,\ldots,m.$$

The joint probability of $(D_1 = d_1,\ldots,D_n = d_n)$, where $(d_1,\ldots,d_n) \in \{0,1,\ldots,m\}^n$, is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{j=1}^{n}[F(x_j - \mu_{d_{j-1}}) - F(x_j - \mu_{d_j})]\, dG_n(x_1,\ldots,x_n). \quad (5.1)$$

As in the binary case the process can be viewed as a truncation of a new process $\{V_n\}_{n=-\infty}^{\infty}$, except that now the truncations occur at the points $(\mu_0,\mu_1,\ldots,\mu_{m-1})$,

and a characterization Lemma analogous to Lemma 3.2 can be proven. The approximations A(1) to A(6) can be extended to the multinomial setting, although they will be given by integral representations rather than closed-form expressions. However, because the integrals are at most 3 dimensional, they can be evaluated quite efficiently.

6. Some Comments on Estimation

In the previous sections probability calculations have involved the parameters of the underlying process and the response function; ordinarily, these parameters are unknown and, therefore, need to be estimated. Prior to estimation, the model to be estimated must be determined. First, the appropriate underlying process and response function are identified, and, second, the parameters involved are estimated. A third step is diagnostic checking; this includes goodness-of-fit tests, residual analysis, and plots to see if the data shows any gross departures from the fitted model. For continuous-valued time series data, Box and Jenkins (1970) have developed procedures for these three steps of analysis. This section is a brief introduction to the model identification and estimation steps for binary processes. The ideas developed here are introductory, and it is hoped that they will serve as a first step towards a more comprehensive theory. This section deals with data from a binary process generated by an underlying Gaussian process and response function.

Let $\{D_n\}_{n=-\infty}^{\infty}$ be a binary process generated by $\{X_n\}_{n=-\infty}^{\infty}$, a Gaussian process with $E(X_n) = 0$, $Var(X_n) = \tau^2$, and correlation structure $\{\rho(j)\}_{j=1}^{\infty}$, and $F = \Phi_{0,b^2}$ = c.d.f. of $N(0,b^2)$. The autocorrelation function $\{D_n\}_{n=-\infty}^{\infty}$ is given by equation (3.7):

$$\delta(s) = \frac{2 \arcsin [\rho(s)/(1+k)]}{\pi} .$$

Given data $D_1 = d_1$, $D_2 = d_2, \ldots, D_n = d_n$, if $\delta(s)$ can be estimated from the sample, say, by $\tilde{\delta}_n(s)$, then $[\frac{\rho(s)}{1+k}]$ can be estimated (method of moments) inverting equation (1.7) to obtain

$$\frac{\rho(s)}{1+K} = \sin\left[\frac{\Pi}{2}\tilde{\delta}(s)\right] \quad .$$

As for probabilistic statements concerning the binary process, $\{D_n\}_{n=-\infty}^{\infty}$, $\{\frac{\rho(s)}{1+K}\}_{s=1}^{\infty}$, rather than $\{\rho(s)\}_{s=1}^{\infty}$, are the appropriate parameters. One estimate of $\delta(s)$ is the sample autocorrelation of lag s, $\delta_n(s)$:

$$\hat{\delta}_n(s) = \frac{1}{n-s}\sum_{i=1}^{n-s}(-1)^{d_i + d_{i+s}} \quad .$$

For s = 1, $\hat{\delta}_n(1)$ counts the number of changes of sign (i.e., directional changes) in the $(d_1, d_2, \ldots, d_n)$, relative to the number possible, n-1. Kedem (1980b) and Lomnicki and Zaremba (1955) develop some asymptotic results for this and related estimators. These estimates do not involve assumptions about the form of the underlying autocorrelation structure $\{\rho(s)\}_{s=1}^{\infty}$. That is, $\{X_n\}_{n=-\infty}^{\infty}$ need not have a specified ARMA (p,q) form. However, the model assumes that the response function is symmetrical, that is, $F = \Phi_{\mu 0, b^2}$ and $\mu_0 = 0$. Without constraints on the correlation structure, $k = b^2/\tau^2$, the ratio of the response function variance to that of the underlying process cannot be estimated. If we assume a specific ARMA model for $\{X_n\}_{n=-\infty}^{\infty}$, then as a result of the constraints that knowing p and q impose on the autocorrelations, k can be estimated. The variance of $\hat{\delta}_n(s)$ is

$$\mathrm{Var}(\hat{\delta}_n(s)) = \left(\frac{1}{n-s}\right)^2 \sum_{i=1}^{n-s}\sum_{j=1}^{n-s} E\left((-1)^{D_i + D_j + D_{i+s} + D_{j+s}}\right) - [\delta(s)]^2$$

and can be evaluated if 4-dimensional probabilities can be determined. Plackett (1954), McFadden (1960), and Gehrlein (1980), among others, discuss numerical a approximations to these probabilities. However, the appropriate approximation depends on the magnitudes of the correlations. Lomnicki and Zaremba (1955) give a bound for these 4-dimensional probabilities.

7. Concluding Remarks and Future Directions

Data which is binary, recorded sequentially in time, and correlated is common. In Section 2 a family of models for such data which allow for arbitrary correlation structure was proposed. The memory of such models dies out in a continuous manner. It is assumed that an underlying process with continuous-valued state space generates the discrete process through a response function F. An important question is, how much of the structure of the underlying process passes through to the discrete process? The structure with which we are concerned is that of being Markovian. If the states of a Markov chain are lumped together to create a smaller number of new states, the question becomes whether or not the new chain is Markovian (i.e., has the property of lumpability) (Kemeny and Snell (1960), pp. 123-140, and Burke and Rosenblatt (1958)). Therefore, the above question can be viewed as a stochastic version of the lumping of a Markov process into two states, zero and one.

A future direction is the measurement of the loss of the Markov property as a result of discretization. In Section 3 it was shown that the discretization could be viewed as the addition of i.i.d. noise followed by truncation; each of these distorts the Markov property. In a forthcoming paper a measure of the loss of the Markov property due to discretization is proposed. In that paper, the loss is decomposed into a loss due to the noise and a loss due to truncation. The relationship of the above models to those other areas--functions of Markov chains, source-coding, and quantization in signal processing--will be

explored. In Section 5 one aspect of the above question, the ease of forecasting from the discrete process when the underlying process is Markovian, was considered. It was shown that although the discrete process is not Markovian, for prediction purposes it is reasonable to act as if it is. Bounds on and approximations to the Probabilities of Error for the optimal predictor were derived.

REFERENCES

Abrahamson, I. G. (1964). Orthant Probabilities for the Quadrivariate Normal Distribution. Ann. Math. Statist. 35, pp. 1685-1703.

Billingsley, P. (1961). Statistical Inference for Markov Processes. University of Chicago Press, Chicago.

Box, G.E.P., and Jenkins, G.M. (1970). Time Series Analysis, Forecasting, and Control. Holden-Day, San Francisco.

Burke, C. J., and Rosenblatt, M. (1958). A Markovian Function of a Markov Chain. Ann. Math. Statist. 29, pp. 1112-1122.

Bush, R. R., and Mosteller, F. (1951). A Mathematical Model for Simple Learning. In Luce, R. D.; Bush, R. R.; and Galanter, E. (Eds.), Readings in Mathematical Psychology, (1963). John Wiley and Sons, Inc., New York. Vol. 1, pp. 289-299.

Cheng, M. C. (1969). The Orthant Probabilities of Four Gaussian Variates. Ann. Math. Statist. 40, pp. 152-161.

Choi, J. R. (1975). An Equality Involving Orthant Probabilities. Comm. Statist. 4, No. 12, 1167-1175.

Das Gupta, Eaton, Olkin, Perlman, Savage, and Sobel (1972). Inequalities on the Probabilities Content of Convex Regions for Elliptically Contoured Distributions. Proc. Sixth Berkeley Symp. Math. Statist. Prob. Berkeley and Los Angeles, University of California Press, Vol. 2, pp. 241-265.

David, F. N. (1953). A Note on the Evolution of the Multivariate Normal Integral. Biometrika 40, pp. 458-459.

Downing, D. J.; Pike, D. H.; and Morrison, G. W. (1980). Application of the Kalman Filter To Inventory Control. Technometrics 22, No. 1, pp. 17-22.

Duncan, D. B., and Horn, S. D. (1972). Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis. J. Amer. Statist. Assoc. 67, pp. 815-821.

Feller, W. (1971). An Introduction to Probability Theory and its Applications. Vol. II, Second Edition, John Wiley & Sons, Inc., New York.

Gehrlein, W. V. (1979). A Representation for Quadrivariate Normal Positive Orthant Probabilities. Comm. Statist. B. 8, pp. 349-358.

Granger, C. W. J., and Morris, M. (1976). Time Series Modelling and Interpretation. J. R. Statist. Soc. A. 38, pp. 246-257.

Gupta, S. S. (1963). Probability Integrals of Multivariate Normal and Multivariate t. Ann. Math. Statist. 34, pp. 792-828.

Heckman, J. J. (1979). Statistical Models for Discrete Panel Data Report 7902, Center for Mathematical Studies in Business and Economics, University of Chicago, January 1979.

Jacobs, P. A., and Lewis, P. A. W. (1978). Discrete Time Series Generated by Mixtures. I. Correlation And Runs Properties. J. R. Statist. Soc. B. 40, No. 1, pp. 94-105.

_____ (1978). Discrete Time Series Generated by Mixtures. II. Asymptotic Properties. J. R. Statist. Soc. B. 40, No. 2, pp. 222-228.

Jazwinski, A. H. (1970). Stochastic Processes and Filtering Theory. Academic Press, New York.

Kedem, B. (1980). Binary Time Series. Marcel Dekker, Inc., New York.

_____ (1980). Estimation of the Parameters in Stationary Autoregressive Process After Hard Limiting. J. Ann. Statist. Assoc. 75, pp. 146-153.

Keenan, D. M. (1980). Time Series Analysis of Binary Data. Technical Report No. 130, Department of Statistics, University of Chicago.

Kemeny, J. G., and Snell, J. L. (1960). Finite Markov Chains. D. Van Nostrand Company, Inc., Princeton, New Jersey.

Lamperti, J., and Suppes, P. (1959). Chains of Infinite Order and Their Application to Learning Theory. Pacific J. Math. 9, pp. 739-754.

Larson, R. (1979). The Significance of Solitude in Adolescents' Lives. Ph.D. Dissertation, Committee on Human Development, University of Chicago.

Lomnicki, Z. A., and Zaremba, S. K. (1955). Some Applications of Zero-One Processes. J. R. Statist. Soc. B. 17, pp. 243-255.

McFadden, J. A. (1960). Two Expansions for the Quadrivariate Normal Integral. Biometrika. 47, 3 and 4, pp. 325-333.

Moran, P. A. P. (1956). The Numerical Evaluation of a Class of Integrals. Proc. Cambridge Phil. Soc. 52, pp. 230-233.

Plackett, R. L. (1954). A Reduction Formula for Normal Multivariate Integrals. Biometrika. 41, pp. 351-360.

Schlafli, L. (1858). On the Multiple Integral $\&^n d_x, d_y, \ldots d_z$ whose limits are $p_1 = a_1 x + b_1 y + \ldots + h_1 z_1$ $p_2 > 0, \ldots, p_n > 0$ and $x^2 + \ldots + z^2$ Quart. J. Math. 2, pp. 269-301, 3, 54-68, and 97-108.

Slepian, D. (1962). The One-Sided Barrier Problem for Gaussian Noise. Bell System Tech. J. 41, pp. 463-501.

Sondhi, M. M. (1961). A Note on the Quadrivariate Normal Integral. Biometrika 48, pp. 201-203.