AN APPROACH TO TIME SERIES ANALYSIS[1]

By Emanuel Parzen

Stanford University

Summary. It may fairly be said that modern time series analysis is a subject which embraces three fields which while closely related have tended to develop somewhat independently. These fields are (i) statistical communication and control theory, (ii) the probabilistic (and Hilbert space) theory of stochastic processes possessing finite second moments, and (iii) the statistical theory of regression analysis, correlation analysis, and spectral (or harmonic) analysis of time series. In this paper it is my aim to show the close relation between these fields and to summarize some recent developments.

The topics discussed are (i) stationary time series and their statistical analysis, (ii) prediction theory and the Hilbert space spanned by a time series, and (iii) regression analysis of time series with known covariance function. In particular, I describe a new approach to prediction and regression problems using reproducing kernel Hilbert spaces.

1. Introduction. A set of observations arranged chronologically is called a time series. Time series are observed in connection with quite diverse phenomena, and by a wide variety of researchers, such as (1) the economist observing yearly wheat prices, (2) the geneticist observing daily egg production of a certain breed of hen, (3) the meteorologist studying daily rainfall in a given city, (4) the physicist studying the ambient noise level at a given point in the ocean, (5) the aerodynamicist studying atmospheric turbulence gust velocities, (6) the electronic engineer studying the internal noise of a radio receiver, and so on.

Time series analysis constitutes one of the most important tools of the economist. Consider the prices or quantities of commodities traded on an exchange. The record of prices or quantities over time may be represented as a fluctuating function (or wiggly record). The analysis of such economic time series is a problem of great interest to economists desiring to explain the dynamics of economic systems and to speculators desiring to forecast prices.

Techniques of time series analysis have long been used in science and engineering (for example, to smooth data and to search for "periodicities" [6]). The theory and practice of time series analysis is assuming new importance in the space age since a wide variety of problems involving communication and/or control (involving such diverse problems as the automatic tracking of moving

Received September 27, 1960.

[1] An address presented on August 25, 1960 at the Stanford meetings of the Institute of Mathematical Statistics by invitation of the IMS Committee on Special Invited Papers. This paper was prepared with the partial support of the Office of Naval Research (Nonr-225-21).

objects, the reception of radio signals in the presence of natural and artificial disturbances, the reproduction of sound and images, the design of guidance systems, the design of control systems for industrial processes, and the analysis of any kind of record representing observation over time) may be regarded as problems in time series analysis.

To represent a time series, one proceeds as follows. The set of time points at which measurements are made is called T. The observation made at time t is denoted by $X(t)$. The set of observations $\{X(t), t \varepsilon T\}$ is called a time series.

In regard to the index set T, there are cases of particular importance. One may be observing (i) a discrete parameter time series $X(t)$, in which case one assumes T is a finite set of points written $T = \{1, 2, \cdots, N\}$, (ii) a continuous parameter time series, in which case T is a finite interval written $T = \{t: 0 \leqq t \leqq L\}$, (iii) a multiple (discrete or continuous parameter) time series $\{(X_1(t), \cdots, X_k(t)), t \varepsilon T'\}$ which may be written as a time series $\{(X(t), t \varepsilon T\}$ with index set $T = \{(j, t): j = 1, \cdots, k \text{ and } t \varepsilon T'\}$, or (iv) a space field $X(x, y, z, t)$ defined on space-time which is a function of three coordinates of position and one coordinate of time.

The basic idea of the statistical theory of analysis of a time series $\{X(t), t \varepsilon T\}$ is to regard the time series as being an observation made on a family of random variables $\{X(t), t \varepsilon T\}$; that is, for each t in T, $X(t)$ is an observed value of a random variable. A family of random variables $\{X(t), t \varepsilon T\}$ is called a *stochastic process*. An observed time series $\{X(t), t \varepsilon T\}$ is thus regarded as an observation (or, in a different terminology, a realization) of a stochastic process $\{X(t), t \varepsilon T\}$.

It has been pointed out by various writers (see, for example, Neyman [34]) that there are two broad categories of statistical problems: problems of stochastic model building for natural phenomena and problems of statistical decision making. These two categories of problems are well illustrated in the analysis of economic time series; some study time series in order to understand the mechanism of the economic system while others study time series with the simple aim of being able to forecast, for example, stock market prices. In general, it may be said that the aims of time series analysis are

(1) to understand the mechanism generating the time series,

(2) to predict the behavior of the time series in the future. To attack either of these problems, one adopts a model for the time series.

A model often adopted for the analysis of an observed time series $\{X(t), t \varepsilon T\}$ is to regard $X(t)$ as the sum of two functions:

$$X(t) = m(t) + Y(t), \qquad t \varepsilon T. \tag{1.1}$$

We call $m(t)$ the mean value function and $Y(t)$ the fluctuation function.

The stochastic process $Y(t)$ is assumed to possess finite second moments, and to have zero means and *covariance kernel*

$$K(s, t) = E[Y(s)Y(t)]. \tag{1.2}$$

In addition it is often assumed that $Y(t)$ is a *normal process* in the sense that

for every finite subset $\{t_1, \cdots, t_n\}$ of T, the random variables $Y(t_1), \cdots, Y(t_n)$ are jointly normally distributed.

The mean value function

$$m(t) = E[X(t)] \tag{1.3}$$

is assumed to belong to a known class M of functions. For example, M may be the set of all linear combinations of q known functions $w_1(t), \cdots, w_q(t)$; then, for t in T,

$$m(t) = \beta_1 w_1(t) + \cdots + \beta_q w_q(t), \tag{1.4}$$

for some coefficients $\beta_1, \cdots, \beta_q$ to be estimated. Other possible assumptions often made concerning the mean value function $m(t)$ are as follows: (i) $m(t)$ represents a *systematic oscillation*,

$$m(t) = \sum_{j=1}^{q} A_j \cos(\omega_j t + \varphi_j) \tag{1.5}$$

in which the amplitudes A_j, the angular frequencies ω_j, and the phases φ_j are constants, some of which are given and the rest are unknown and are to be estimated; (ii) $m(t)$ represents a polynomial trend,

$$m(t) = \sum_{j=0}^{q-1} \beta_j t^j, \tag{1.6}$$

an assumption often adopted if $m(t)$ represents the trajectory [given by $m(t) = x_0 + vt + \frac{1}{2}at^2$, say] of a moving object, or (iii) $m(t)$ is the sum of a systematic oscillation and a polynomial trend, an assumption traditionally adopted in treating economic time series.

Early workers in time series analysis sought to explain the dependence between successive observations of a time series $X(t)$ by assuming that $X(t)$ [sometimes written X_t] was generated by a scheme of the following kind:

$$X_t = m(t) + Y_t \tag{1.7}$$

where $m(t)$ represents a systematic oscillation of the form of (1.5) and the fluctuations $Y_1, \cdots, Y_n$ are assumed to be independent, normal random variables with mean 0 and common unknown variance σ^2.

The model given by (1.7) is called the scheme of *hidden periodicities* and was first introduced by Schuster ([47], [48]). The method used to estimate the frequencies ω_j (or, equivalently, the periods $2\pi/\omega_j$) is called *periodogram analysis*. The problem of tests of significance in periodogram analysis ([11], [18]) played an important role in the early history of time series analysis.

Approaches to time series analysis which seem to be more fruitful than periodogram analysis (see Kendall [22]) were pioneered by Yule and Slutsky in the 1920's.

Yule's researches [60] led to the notion of the autoregressive scheme, in which a time series X_t is assumed to be generated as a linear function of its past values,

plus a random shock; in symbols, for some integer m (called the order of the autoregressive scheme) and constants $a_1, \cdots, a_m$

$$X_t = a_1 X_{t-1} + \cdots + a_m X_{t-m} + \eta_t$$

in which the sequence $\{\eta_t\}$ consists of independent identically distributed random variables. In particular, Yule showed that an autoregressive scheme of order 2 provided a better model for sunspots than did the scheme of hidden periodicities.

Slutsky's researches [51] led to the notion of a *moving average scheme*, in which a time series X_t is assumed to be generated as a finite moving average of a sequence of independent and identically distributed random variables $\{\eta_t\}$; in symbols, for some integer m and constants $a_0, \cdots, a_m$

$$X_t = a_0 \eta_t + a_1 \eta_{t-1} + \cdots + a_m \eta_{t-m}.$$

Slutsky showed that moving averages exhibit properties of disturbed periodicity and consequently can be used as a model for oscillatory time series. In particular, Slutsky proved the Sinusoidal Limit Theorem which showed that a sine wave could be approximated by a moving average scheme.

In the 1930's and 1940's, the probabilistic theory of stationary time series was developed, first as a result of the development of ergodic theory and then as a result of prediction theory. That the autoregressive and moving average schemes may be interpreted as special cases of the theory of stationary processes was pointed out by Wold [58] in 1938 (see [57], p. 169). Thus the link was established between the statistical theory of analysis of time series and the probabilistic theory of the structure of time series. In the last twenty years, an extensive literature has developed exploring this link.

It may fairly be said that modern time series analysis is a subject which embraces three fields which while closely related have tended to develop somewhat independently. These fields are (i) statistical communication and control theory ([26], [32]) (ii) the probabilistic (and Hilbert space) theory of stochastic processes possessing finite second moments ([9], Chaps. 9–12; [28], Chap. 10), and (iii) the statistical theory of regression analysis, correlation analysis, and spectral (or harmonic) analysis of time series ([13], [17], [23], [58]). In this paper it is my aim to show the close relation between these fields and to summarize some recent developments with which I have been closely associated. The contents of the paper are as follows.

(I) *Stationary time series and their statistical analysis.* While it is a fiction to regard an observed time series as having zero means, it is mathematically convenient to consider the analysis of time series under this assumption. Consequently, one may consider the analysis of an observed time series $\{X(t), t \varepsilon T\}$ with vanishing mean value function and *unknown* covariance function.

It has long been traditional among physical scientists to regard time series as arising from a superposition of sinusoidal waves of various amplitudes, frequencies, and phases. In the theory of time series analysis and statistical communications theory, a central role is played by the notion of the spectrum of a

time series. For the time series analyst, the spectrum represents a basic tool for determining the mechanism generating an observed time series. For the communication theorist, the spectrum provides the major concept in terms of which to analyze the effect of passing stochastic processes (representing either signal or noise) through linear (and, to some extent, non-linear) devices. *Spectral* (or *harmonic*) *analysis* is concerned with the theory of the decomposition of a time series into sinusoidal components. For many time functions $\{X(t), -\infty < t < \infty\}$, such a decomposition is provided by the Fourier transform

$$S(\omega) = \int_{-\infty}^{\infty} e^{-it\omega} X(t)\, dt.$$

Unfortunately, no meaning can be attached to this integral for many stochastic processes $\{X(t), -\infty < t < \infty\}$ since their *sample functions are nonperiodic undamped functions* and therefore do not belong to the class of functions dealt within the usual theories of Fourier series and Fourier integrals. Nevertheless, it is possible to define a notion of *harmonic analysis of stochastic processes* (that is, a method of assigning to each frequency ω a measure of its contribution to the "content" of the process) as was first shown by N. Wiener [56] and A. Khintchine [24]. Among the stochastic processes which possess a harmonic analysis, *stationary* processes are most important since a time series may be represented as a superposition of sinusoidal waveforms with "independent amplitudes" if and only if it is stationary (see Section 4 for a more precise form of this assertion). In Section 2, some basic results concerning stationary processes are summarized.

Much of the recent statistical literature on time series analysis has been concerned with questions of statistical inference on stationary time series and especially with

(i) deriving the exact and asymptotic distributions of various estimates of the covariance functions $R(v)$ and the normalized covariance (or correlation) function $\rho(v) = R(v)/R(0)$ of a stationary time series,

(ii) fitting stationary time series by mechanisms (such as autoregressive schemes or moving average schemes) which are completely specified except for a finite number of parameters, and with estimating the parameters of such schemes,

(iii) estimating (and forming confidence sets) for the spectral density function and spectral distribution function of a stationary time series.

For many purposes, it is preferable to estimate the spectrum of a stationary time series rather than its correlation function, since many aspects of a stationary time series are best understood in terms of its spectrum. The spectrum enables one to (i) investigate the physical mechanism generating a time series, (ii) determine the behavior of a dynamic linear system in response to random excitations, and (iii) possibly simulate a time series. Other uses of the spectrum are as operational means (i) of transmitting or detecting signals, (ii) of classifying records of phenomena such as brain waves, (iii) of studying radio propagation

phenomena, and (iv) of determining characteristics of control systems. The theory of statistical spectral analysis is too extensive to be reviewed here. For surveys of this theory, see Bartlett [3], Hannan [17], Blackman and Tukey [4], Jenkins [20], Parzen [41], Rosenblatt [45], and Tukey [53].

A comprehensive survey of the results available on topics (i) and (ii) has been given recently by Hannan [17]. Other comprehensive reviews are given by Bartlett [3], Moran [33], and Wold [57], Chap. 11; in these reviews one may find references to the work of M. S. Bartlett, H. E. Daniels, J. Durbin, E. J. Hannan, G. M. Jenkins, M. H. Quenouille, A. M. Walker, G. S. Watson, and P. Whittle. In Section 8, I have attempted to give an introduction to some of the large sample results available on topics (i) and (ii).

(II) *Prediction theory and the Hilbert space spanned by a time series.* A basic problem in time series analysis is that of *minimum mean square error linear prediction.* Let Z be an unobserved random variable with finite second moment. Let $\{X(t), t \varepsilon T\}$ be an observed time series. One seeks that random variable, linear in the observations, whose mean square distance from Z is smallest. In other words, if one desires to predict the value of Z, on the basis of having observed the values of the time series $\{X(t), t \varepsilon T\}$, one method might be to take that linear functional in the observations, denoted by $E^*[Z \mid X(t), t \varepsilon T]$, whose mean square error as a predictor is least. (The symbol E^* is used to denote a predictor because in the case of jointly normally distributed random variables, the best linear predictor $E^*[Z \mid X(t), t \varepsilon T]$ coincides with the conditional expectation $E[Z \mid X(t), t \varepsilon T]$; for an elementary discussion of this fact, see Parzen ([38], p. 387). Indeed, it should be noted that in any event the conditional expectation $E[Z \mid X(t), t \varepsilon T]$ can be defined as the minimum mean square error *non-linear* predictor.)

The prediction problem has provided a framework in terms of which many problems of statistical communication theory have come to be formulated. The pioneering work on prediction theory was done by Wiener [56a] and Kolmogorov [25] who were concerned with a stationary time series which had been observed over a semi-infinite interval of time. They sought predictors which had minimum mean square over all possible linear predictors. Wiener showed how the solution of the prediction problem could be reduced to the solution of the so-called Wiener-Hopf integral equation, and gave a method (spectral factorization) for the solution of the integral equation. Simplified methods of solution of this equation in the practically important special case of rational spectral density functions were given by Zadeh and Ragazzini [61] and Bode and Shannon [5]. Zadeh and Ragazzini [62] also treated the problem of regression analysis of time series with stationary fluctuation function, by reducing the problem to one involving the solution of a Wiener-Hopf equation. There then developed an extensive literature, seeking to treat prediction and smoothing problems involving a finite time of observation and non-stationary time series. The methods employed were either to reduce the problem to the solution of a suitable integral equation (generalization of the Wiener-Hopf equation) or to employ expansions (in a series of suitable eigen functions) of the time series involved.

As a result of these developments, prediction theory has turned out to provide a theory of the structure of time series and to provide mathematical tools for the solutions of other problems besides the prediction problem, especially regression problems. In Section 3, it is shown how the prediction problem leads naturally to the introduction of the important notion of the *Hilbert space spanned by a time series* which plays a central role in modern time series analysis. In Sections 4 and 6, I describe an approach to prediction and regression problems (in terms of reproducing kernel Hilbert spaces) which may be called coordinate free, and which by the introduction of suitable coordinate systems contains previous approaches as special cases.

The approach I take seems to me to be a rigorous version of an approach that is being developed in the Soviet Union by V. S. Pugachev [44]. Pugachev has in recent years advanced a point of view, which he calls the method of canonic representations of random functions, for which in one of his articles [43] he makes the following claim. "The results of this article, together with the results of [previous] papers, permits us to state that the method of canonic representations of random functions is the foundation of the modern statistical theory of optimum systems." It is my feeling that reproducing kernel Hilbert spaces provide a more powerful and more elegant means of achieving in a unified manner the results which Pugachev has sought to unify by the method of canonic representations.

(III) *Regression analysis of time series with known covariance function:* Let the observed time series be of the form of (1.1) with unknown mean value function $m(t)$ and known covariance function $K(s, t)$. Various methods of forming estimates of $m(t)$ are available. The most important methods are classical least squares estimation and minimum variance linear unbiased estimation. In the case of normally distributed observations, one has in addition the methods of maximum likelihood estimation and minimum variance unbiased estimation. In Sections 6 and 7, it is shown how Hilbert space techniques may be used to form explicit expressions for these estimates in terms of certain so-called reproducing kernel inner products.

There are, of course, large numbers of important problem areas of time series analysis which have not been mentioned in the foregoing such as (i) the problem of the distribution of zero-crossings and extrema of a time series (see references, see Longuet-Higgins [30] and Slepian [50]), (ii) the problem of the asymptotic efficiency of various classes of estimates of regression coefficients (see Grenander and Rosenblatt [13], Hebbe [19], and Striebel [52]), (iii) the use of filters to eliminate or extract trend or other components of a time series, and (iv) the distribution of various functionals of a time series, such as quadratic forms.

Further, the statistical analysis of multiple time series is not discussed. The relations that exist between different time series is on the whole a problem of greater interest than the relations that exist within a single time series. The results which exist under categories (I), (II), and (III) for univariate time series can be formally extended to multiple time series. However, many new problems arise which have not been thoroughly investigated.

A word should be said about the references given at the end of this paper. I have given a representative list rather than a complete list. Fortunately, a complete list of references will soon be available. The International Statistical Institute is compiling a bibliography on *Time Series and Stochastic Processes* which is to list and classify books and papers published, in the years 1900–1959, on both theory and applications. A bibliography (Parzen [42a]) of American publications has been compiled at Stanford for inclusion in the I.S.I. bibliography; a limited number of copies of this bibliography are available, and may be obtained by writing to the author. A bibliography is also given by Deming [8].

2. Stationary time series. A discrete parameter time series

$$\{X(t), t = 0, \pm 1, \cdots\}$$

or a continuous parameter time series $\{X(t), -\infty < t < \infty\}$ is said to be (weakly or wide-sense) stationary if the product moment

$$E[X(s)X(s + t)] = R(t)$$

is a function only of t. One calls $R(\cdot)$ the covariance function of the stationary time series.

It was shown by Khintchine [24] in 1933 that in the continuous parameter case there exists a non-decreasing bounded function $F(\omega)$, defined for $-\infty < \omega < \infty$, such that

$$R(t) = \int_{-\infty}^{\infty} e^{it\omega}\, dF(\omega), \qquad -\infty < t < \infty, \tag{2.1}$$

if it is assumed that $R(\cdot)$ is continuous at $t = 0$. Wold [58] in 1938 showed that in the discrete parameter case there exists a non-decreasing bounded function $F(\omega)$, defined for $-\pi \leqq \omega \leqq \pi$, such that

$$R(t) = \int_{-\pi}^{\pi} e^{it\omega}\, dF(\omega), \qquad t = 0, \pm 1, \cdots. \tag{2.2}$$

The function $F(\omega)$ is called the *spectral distribution function* of the time series. Like a probability distribution function, $F(\omega)$ can be uniquely written as the sum,

$$F(\omega) = F_d(\omega) + F_{sc}(\omega) + F_{ac}(\omega),$$

of three distribution functions with the following properties. The function $F_{ac}(\omega)$ is absolutely continuous and is the integral of a non-negative function $f(\omega)$ called the *spectral density function* of the time series. The function $F_d(\omega)$ is a purely discontinuous (or discrete or step) function:

$$F_d(\omega) = \sum_{\omega_j \leqq \omega} \Delta F(\omega_j)$$

where $\{\omega_j\}$ are the discontinuity points of $F(\omega)$, and $\Delta F(\omega) = F(\omega + 0) - F(\omega - 0)$. Finally, $F_{sc}(\omega)$ is a singular continuous function.

It is usually assumed that physically observed time series have a spectral distribution function of the following form:

$$F(\omega) = \sum_{\substack{\omega' \text{ such} \\ \text{that } \Delta F(\omega') > 0 \\ \text{and } \omega' \leqq \omega}} \Delta F(\omega') + \int_{-\infty}^{\omega} f(\omega')\, d\omega'$$

where (i) the spectral density function $f(\omega)$ has the property that it is an integrable non-negative function which is continuous except at a finite number of points where it has finite left-hand and right-hand limits, and (ii) the set of frequencies at which the spectral jump function (or spectral mass function) $\Delta F(\omega)$ is positive contains at most countably infinite many points distributed on the real line in such a way that in any finite interval there are only a finite number of points of positive spectral mass. If these conditions are satisfied, we say that the time series has a *mixed spectrum*. If the spectral density function vanishes for all ω, we say that the time series has a *discrete* spectrum. If the spectral jump function $\Delta F(\omega)$ vanishes for all ω, we say that the time series has a *continuous* spectrum.

In terms of the spectral distribution function one can characterize various representations (or models) for a stationary time series $X(t)$. For example, it may be shown that a discrete parameter time series with a mixed spectrum whose spectral density function satisfies the condition

$$\int_{-\pi}^{\pi} \log f(\omega)\, d\omega > -\infty$$

may be written

$$X(t) = \sum_{\nu} A_\nu e^{it\omega_\nu} + \sum_{\nu=0}^{\infty} c_\nu \eta(t - \nu) \tag{2.3}$$

for suitable sequences of frequencies $\{\omega_\nu\}$, constants $\{c_\nu\}$, and uncorrelated random variables $\{A_\nu\}$ and $\{\eta_\nu\}$. In view of (2.3) one sees that the scheme of hidden periodicities and the scheme of moving averages may be viewed as a special kind of stationary process. Similarly, it may be shown that an autoregressive scheme (where the η_t are uncorrelated rather than independent) corresponds to a stationary time series whose distribution function is absolutely continuous and whose spectral density function is of the form

$$f(\omega) = \left[2\pi \left| \sum_{k=0}^{m} b_k e^{ik\omega} \right|^2 \right]^{-1}$$

for suitable constants $b_0, \cdots, b_m$. To prove these assertions, one uses the Hilbert space representation theory described in Section 4.

3. The problem of minimum mean square error linear prediction. In order to show existence and uniqueness, and to obtain conditions characterizing, the best

linear predictor, we need to introduce the notion of a Hilbert space. (For a discussion of Hilbert space theory see any suitable text, such as Halmos [16].)

DEFINITION 3A. By an abstract Hilbert space is meant a set H whose members $u, v, \cdots$ are usually called vectors or points which possesses the following properties.

(I) H is a linear space [that is, for any vectors u and v in H, and real number a, there exist vectors, denoted by $u + v$ and au respectively, which satisfy the usual algebraic properties of addition and multiplication; also there exists a zero vector 0 with the usual properties under addition].

(II) H is an inner product space [that is, to every pair of points u and v in H there corresponds a real number, written (u, v) and called the inner product of u and v, possessing the following properties: for all points u, v, and w in H, and every real number a,

(i) $(au, v) = a(u, v)$

(ii) $(u + v, w) = (u, w) + (v, w)$

(iii) $(v, u) = (u, v)$

(iv) $(u, u) > 0$ if and only if $u \neq 0$].

(III) H is a complete metric space under the norm $\|u\| = (u, u)^{\frac{1}{2}}$ [that is, if $\{u_n\}$ is a sequence of points such that $\|u_m - u_n\| \to 0$ as $m, n \to \infty$ then there is a vector u in H such that $\|u_n - u\|^2 \to 0$ as $n \to \infty$].

In order to define the notion of the Hilbert space spanned by a time series, we first define the notion of the Hilbert space spanned by a family of vectors.

DEFINITION 3B. Let T be an index set, and let $\{u(t), t \varepsilon T\}$ be a family of members of a Hilbert space H. The linear manifold spanned by the family $\{u(t), t \varepsilon T\}$, denoted $L(u(t), t \varepsilon T)$, is defined to be the set, consisting of all vectors u in H which may be represented in the form $u = \sum_{i=1}^{n} c_i u(t_i)$ for some integer n, some constants $c_1, \cdots, c_n$, and some points $t_1, \cdots, t_n$ in T. The Hilbert space spanned by the family $\{u(t), t \varepsilon T\}$, denoted $V(u(t), t \varepsilon T)$ [or $L_2(u(t), t \varepsilon T)$ if H is the space of square integrable functions on some measure space], is defined to be the set of vectors which either belong to the linear manifold $L(u(t), t \varepsilon T)$ or may be represented as a limit of vectors in $L(u(t), t \varepsilon T)$. If $V(u(t), t \varepsilon T)$ coincides with H, we say that $\{u(t), t \varepsilon T\}$ *spans* H.

DEFINITION 3C. The Hilbert space spanned by a time series $\{X(t), t \varepsilon T\}$, denoted by $L_2(X(t), t \varepsilon T)$, is defined to consist of all random variables U which are either finite linear combinations of the random variables $\{X(t), t \varepsilon T\}$ or are limits of such finite linear combinations in the norm corresponding to the inner product defined on the space L_2 of square integrable random variables by

$$(U, V) = E[UV]. \tag{3.1}$$

In words, $L_2(X(t), t \varepsilon T)$ consists of all linear functionals in the time series.

We next state without proof the projection theorem for an abstract Hilbert space.

PROJECTION THEOREM. *Let H be an abstract Hilbert space, let M be a Hilbert subspace of H, let v be a vector in H, and let v^* be a vector in M. A necessary and*

sufficient condition that v^* is the unique vector in M satisfying

$$\|v^* - v\| = \min_{u \text{ in } M} \|u - v\| \tag{3.2}$$

is that

$$(v^*, u) = (v, u) \qquad \textit{for every } u \textit{ in } M. \tag{3.3}$$

The vector v^* satisfying (3.2) *is called the projection of v onto M, and will here be written $E^*[v \mid M]$.*

In the case that M is the Hilbert space spanned by a family of vectors $\{x(t), t \varepsilon T\}$ in H, we write $E^*[v \mid x(t), t \varepsilon T]$ to denote the projection of v onto M. In this case, a necessary and sufficient condition that v^* satisfy (3.3) is that

$$(v^*, x(s)) = (v, x(s)) \qquad \text{for every } s \text{ in } T. \tag{3.4}$$

We are now in a position to solve the problem of obtaining an explicit expression for the minimum mean square error linear prediction $E^*[Z \mid X(t), t \varepsilon T]$. From (3.4), with H equal to the Hilbert space L_2 of all square integrable random variables, and $v = Z$, it follows that the optimum linear predictor is the unique random variable in $L_2(X(t), t \varepsilon T)$ satisfying, for all s in T,

$$E[E^*[Z \mid X(t), t \varepsilon T]X(s)] = E[ZX(s)]. \tag{3.5}$$

Equation (3.5) may look more familiar if we consider the special case of an interval $T = \{t: a \leqq t \leqq b\}$. If one writes heuristically

$$\int_a^b X(t)w(t)\, dt \tag{3.6}$$

to represent a random variable in $L_2(X(t), t \varepsilon T)$, then (3.5) states that the weighting function $w^*(t)$ of the best linear predictor

$$E^*[Z \mid X(t), t \varepsilon T] = \int_a^b w^*(t)X(t)\, dt, \tag{3.7}$$

must satisfy the generalized Wiener-Hopf equation

$$\int_a^b w^*(t)K(s, t)\, dt = \rho_Z(s), \qquad a \leqq s \leqq b \tag{3.8}$$

where we define

$$K(s, t) = E[X(s)X(t)] \tag{3.9}$$

$$\rho_Z(t) = E[ZX(t)]. \tag{3.10}$$

There is an extensive literature concerning the solution of the integral equation in (3.8); see [39] for references. In my opinion, however, this literature is concerned with an unnecessarily hard problem, as well as one in which the very formulation of the problem makes it difficult to be rigorous. The integral equation in (3.8) possesses a solution only if one interprets $w^*(t)$ as a generalized

function which includes terms which are Dirac delta functions and derivatives of delta functions.

It seems to me that a simple reinterpretation of (3.8) avoids all these difficulties. Let us not regard (3.8) as an integral equation for the weighting function $w^*(t)$. Rather, let us compare (3.7) and (3.8). These equations say that *if one can find a representation for the function* $\rho_Z(s)$ *in terms of linear operations on the functions* $\{K(s, t), t \varepsilon T\}$, *then the minimum mean square error linear predictor* $E^*[Z \mid X(t), t \varepsilon T]$ *can be written in terms of the corresponding linear operations on the time series* $\{X(t), t \varepsilon T\}$. It should be emphasized that the most important linear operations are integration and differentiation. Consequently, the problem of finding the best linear predictor is not one of solving an integral equation, but is one of hunting for a linear representation of $\rho_Z(t)$ in terms of the covariance kernel $K(s, t)$. A general method of finding such representations will be discussed in Sections 4 and 5.

We illustrate the ideas involved by considering a simple example.

EXAMPLE 3A. Consider a stationary time series $X(t)$, with covariance kernel

$$K(s, t) = Ce^{-\beta|t-s|}, \tag{3.11}$$

which one has observed over a finite interval of time, $a \leqq t \leqq b$. Suppose that one desires to predict $X(b + c)$, for $c > 0$. Now, for $a \leqq t \leqq b$,

$$\rho(t) = E[X(t)X(b + c)] = Ce^{-\beta(b+c-t)} = e^{-\beta c}K(b, t). \tag{3.12}$$

In view of (3.12), by the interpretation of (3.7) and (3.8) just stated, it follows that

$$E^*[X(b + c) \mid X(t), a \leqq t \leqq b] = e^{-\beta c}X(b). \tag{3.13}$$

4. Hilbert space representations of time series. In the decade of the 1940's, probabilists began to employ Hilbert space methods to clarify the structure of time series (see [21] and [27]). Among the fundamental theorems proved in this period were the spectral representation theorem for stationary time series, and the Karhunen-Loève representation for random functions of second order on a finite interval. Various workers (especially Grenander [12]) have made use of these representation theorems in treating problems of statistical inference on time series. A representation theorem which does not seem to have found any application is one due to Loève ([27], p. 338) which shows that there is a very intimate connection between time series (random functions of second order) and reproducing kernel Hilbert spaces. It turns out, in my opinion, that reproducing kernel Hilbert spaces are the natural setting in which to solve problems of statistical inference on time series. In this section we define the notion of a Hilbert space representation of a time series and show how this notion may be used to explicitly solve the prediction problem.

The definition we give of the notion of a Hilbert space representation of a time series is based on the following theorem (for proof, see Parzen [37] or [40]).

BASIC CONGRUENCE THEOREM. *Let* H_1 *and* H_2 *be two abstract Hilbert spaces.*

Denote the inner product between two vectors u_1 and u_2 in H_1 by $(u_1, u_2)_1$. Similarly, denote the inner product between two vectors v_1 and v_2 in H_2 by $(v_1, v_2)_2$. Let T be an index set. Let $\{u(t), t \varepsilon T\}$ be a family of vectors which span H_1. Similarly, let $\{v(t), t \varepsilon T\}$ be a family of vectors which span H_2. Suppose that, for every s and t in T,

$$(u(s), u(t))_1 = (v(s), v(t))_2 . \tag{4.1}$$

Then there exists a congruence (a one-one inner product preserving linear mapping) ψ from H_1 onto H_2 which has the property that

$$\psi(u(t)) = v(t), \qquad t \text{ in } T. \tag{4.2}$$

DEFINITION 4A. A family of vectors $\{f(t), t \varepsilon T\}$ in a Hilbert space H is said to be a representation of a time series $\{X(t), t \varepsilon T\}$ if, for every s and t in T,

$$(f(s), f(t))_H = K(s, t) = E[X(s)X(t)]. \tag{4.3}$$

Then there is a congruence (a one-one inner product preserving linear mapping) ψ from $V(f(t), t \varepsilon T)$ onto $L_2(X(t), t \varepsilon T)$ satisfying

$$\psi(f(t)) = X(t) \tag{4.4}$$

and every random variable U in $L_2(X(t), t \varepsilon T)$ may be written

$$U = \psi(g) \tag{4.5}$$

for some unique vector g in $V(f(t), t \varepsilon T)$.

We next show that the representation of a time series as a stochastic integral is best viewed as a Hilbert space representation.

DEFINITION 4B. We call $(Q, \mathbf{B}, \mu)$ a measure space if Q is a set, $\mathbf{B}$ is a σ-field of subsets of Q, and μ is a measure on the measurable space $(Q, \mathbf{B})$. We denote by $L_2(Q, \mathbf{B}, \mu)$ the Hilbert space of all B-measurable real valued functions defined on Q satisfying

$$(f, f)_\mu = \int_Q f^2 \, d\mu < \infty. \tag{4.6}$$

DEFINITION 4C. Let $(Q, \mathbf{B}, \mu)$ be a measure space, and, for every B in $\mathbf{B}$, let $Z(B)$ be a random variable. The family of random variables $\{Z(B), B \varepsilon \mathbf{B}\}$ is called an *orthogonal random set function* with covariance kernel μ if, for any two sets B_1 and B_2 in $\mathbf{B}$,

$$E[Z(B_1)Z(B_2)] = \mu(B_1B_2), \tag{4.7}$$

where, as usual, B_1B_2 denotes the intersection of B_1 and B_2.

The Hilbert space $L_2(Z(B), B \varepsilon \mathbf{B})$ of random variables spanned by an orthogonal random set function may be defined, as was the Hilbert space spanned by a time series, to be the smallest Hilbert subspace of the Hilbert space of all square integrable random variables containing all random variables U of the form $U = \sum_{i=1}^{n} c_i Z(B_i)$ for some integer n, subfamily $\{B_1, \cdots, B_n\} \subset B$, and real

constants $c_1, \cdots, c_n$. On the other hand, $L_2(Q, \mathbf{B}, \mu)$ may be described as the Hilbert space spanned under the norm (4.6) by the family of indicator functions $(I_B, B \varepsilon \mathbf{B})$, where the indicator function I_B of B is defined by $I_B(q) = 1$ or 0 according as $q \varepsilon B$ or $q \varepsilon B$. Now for any B_1 and B_2 in $\mathbf{B}$,

$$(I_{B_1}, I_{B_2})_\mu = \mu(B_1 B_2) = E[Z(B_1)Z(B_2)]. \tag{4.8}$$

Therefore, by the Basic Congruence Theorem, there is a congruence ψ from $L_2(Q, \mathbf{B}, \mu)$ onto $L_2(Z(B), B \varepsilon \mathbf{B})$ such that for any $B \varepsilon \mathbf{B}$,

$$\psi(I_B) = Z(B). \tag{4.9}$$

This fact justifies the following definition of the stochastic integral.

DEFINITION 4D. Let $(Q, \mathbf{B}, \mu)$ be a measure space and let $\{Z(B), B \varepsilon \mathbf{B}\}$ be an orthogonal random set function with covariance kernel μ. For any function f in $L_2(Q, \mathbf{B}, \mu)$ one defines the stochastic integral of f with respect to $\{Z(B), B \varepsilon \mathbf{B}\}$, denoted $\int_Q f\, dZ$, by

$$\int_Q f\, dZ = \psi(f), \tag{4.10}$$

where ψ is the congruence from $L_2(Q, \mathbf{B}, \mu)$ onto $L_2(Z(B), B \varepsilon \mathbf{B})$ determined by (4.9).

We are now in a position to state our version of Karhunen's theorem (see [13], p. 29).

THEOREM 4A. *Let $\{X(t), t \varepsilon T\}$ be a time series with covariance kernel K. Let $\{f(t), t \varepsilon T\}$ be a family of functions in a space $L_2(Q, \mathbf{B}, \mu)$, such that for all s, t in T*

$$K(s, t) = \int_Q f(s)f(t)\, d\mu. \tag{4.11}$$

Then $\{f(t), t \varepsilon T\}$ is a representation for $\{X(t), t \varepsilon T\}$.

If, further, $\{f(t), t \varepsilon T\}$ spans $L_2(Q, \mathbf{B}, \mu)$, then there is an orthogonal random set function $\{Z(B), B \varepsilon \mathbf{B}\}$ with covariance kernel μ such that

$$X(t) = \int_Q f(t)\, dZ, \qquad t \varepsilon T, \tag{4.12}$$

and every random variable U in $L_2(X(t), t \varepsilon T)$ may be represented

$$U = \int_Q g\, dZ \tag{4.13}$$

for some unique function g in $L_2(Q, \mathbf{B}, \mu)$.

PROOF. Let ψ be the congruence from $L_2(f(t), t \varepsilon T)$ onto $L_2(X(t), t \varepsilon T)$ satisfying (4.4). If $\{f(t), t \varepsilon T\}$ spans $L_2(Q, \mathbf{B}, \mu)$, define, for $B \varepsilon \mathbf{B}$, $Z(B) = \psi(I_B)$. It is immediate that $\{Z(B), B \varepsilon \mathbf{B}\}$ is an orthogonal random set function with covariance kernel μ. By the definition of the stochastic integral, (4.12) is merely another way of writing the fact that $X(t) = \psi(f(t))$.

Theorem 4A, together with (2.2) and (2.1), yields the following fundamental result.

SPECTRAL REPRESENTATION THEOREM FOR STATIONARY TIME SERIES. *A discrete parameter time series* $\{X(t), t = 0, \pm 1, \cdots\}$ *is weakly stationary if and only if for some Lebesgue-Stieltjes measures* μ *on the interval* $Q = \{\lambda: -\pi \leqq \lambda \leqq \pi\}$ *the complex exponentials* $\{e^{i\lambda t}, t = 0 \pm 1, \cdots\}$ *form a representation for the time series in* $L_2(Q, \mathbf{B}, \mu)$ *where* $\mathbf{B}$ *is the* σ*-field of Borel subsets of* Q*. Then there exists an orthogonal random set function* $\{Z(B), B \varepsilon \mathbf{B}\}$ *such that*

$$X(t) = \int_{-\pi}^{\pi} e^{it\lambda} Z(d\lambda), t = 0, \pm 1, \cdots . \tag{4.14}$$

A similar theorem holds for continuous parameter time series with

$$Q = \{\lambda: -\infty < \lambda < \infty\}.$$

The representation of a time series as an integral with respect to an orthogonal random set function is not a natural representation, since one may choose such representations of a time series in a multitude of ways. Indeed, if $(Q, \mathbf{B}, \mu)$ is a measure space such that $L_2(X(t), t \varepsilon T)$ and $Q_2(Q, \mathbf{B}, \mu)$ have the same dimension, there are many families $\{f(t), t \varepsilon T\}$ of functions in $L_2(Q, \mathbf{B}, \mu)$ which are a representation for $\{X(t), t \varepsilon T\}$. What one desires is a family $\{f(t), t \varepsilon T\}$ of familiar functions [such as the family of complex exponentials $e^{i\lambda t}$, which are a representation in a suitable space $L_2(Q, \mathbf{B}, \mu)$ for a stationary time series]. I believe there is a natural representation in terms of which to solve problems of statistical inference on time series, namely the representation of a time series with covariance kernel K by the functions $\{K(\cdot, t), t \varepsilon T\}$ in the reproducing kernel Hilbert space $H(K)$.

DEFINITION 4E. A Hilbert space H is said to be a reproducing kernel Hilbert space, with reproducing kernel K, if the members of H are functions on some set T, and if there is a kernel K on $T \otimes T$ having the following two properties; for every t in T (where $K(\cdot, t)$ is the function defined on T, with value at s in T equal to $K(s, t)$):

$$K(\cdot, t) \varepsilon H \tag{4.15}$$

$$(g, K(\cdot, t))_H = g(t) \tag{4.16}$$

for every g in H.

Intuitively, a reproducing kernel Hilbert space is a Hilbert space which contains a function playing the role of the Dirac delta function $\delta(t)$. It should be recalled that, for square integrable functions $f(\cdot)$,

$$\int_{-\infty}^{\infty} f(s)\delta(s - t)\, ds = f(t).$$

Consequently, the kernel $K(s, t) = \delta(s - t)$ satisfies (4.16). However it does not satisfy (4.15), and therefore is not truly a reproducing kernel.

THEOREM 4B (Moore-Aronsjazn-Loève [1], [27]). *The covariance kernel* K *of a time series generates a unique Hilbert space, which we denote by* $H(K)$*, of which* K *is the reproducing kernel.*

Since $K(s,t) = (K(\cdot,s), K(\cdot,t))_{H(K)} = E[X(s)X(t)]$ we immediately obtain the following important theorem.

THEOREM 4C. *Let $\{X(t), t \in T\}$ be a time series with covariance kernel K. Then the family $\{K(\cdot,t), t \in T\}$ of functions in $H(K)$ is a representation for $\{X(t), t \in T\}$. Given a function g in $H(K)$, we denote by $(X, g)_K$ or $(g, X)_K$ the random variable U in $L_2(X(t), t \in T)$ which corresponds to g under the congruence which maps $K(\cdot,t)$ into $X(t)$. We then have the following formal relations: for every t in T, and g, h in $H(K)$,*

$$(X, K(\cdot,t))_K = X(t)$$
$$E[(X,h)_K(X,g)_K] = (h,g)_K \tag{4.17}$$

where we hereafter write $(h,g)_K$ for $(h,g)_{H(K)}$.

The next theorem shows the relationship between the reproducing kernel Hilbert space representation of a time series, and the representation of a time series by an orthogonal decomposition of the form of (4.12).

THEOREM 4D. *Let K be a covariance kernel. If there exist a measure space $(Q, \mathbf{B}, \mu)$, and a family of functions $\{f(t), t \in T\}$ in $L_2(Q, \mathbf{B}, \mu)$ such that (4.11) holds, then the reproducing kernel Hilbert space $H(K)$ corresponding to the covariance kernel K may be described as follows: $H(K)$ consists of all functions g, defined on T, which may be represented in the form*

$$g(t) = \int_Q g^* f(t)\, d\mu \tag{4.18}$$

for some (necessarily unique) function g^* in the Hilbert subspace $L_2(f(t), t \in T)$ of $L_2(Q, \mathbf{B}, \mu)$ spanned by the family of functions $\{f(t), t \in T\}$, with norm given by

$$\|g\|^2 = \int_Q |g^*|^2\, d\mu. \tag{4.19}$$

If $\{f(t), t \in T\}$ spans $L_2(Q, \mathbf{B}, \mu)$, so that $X(t)$ has an orthogonal decomposition (4.12), then we may write

$$(X, g)_K = \int_Q g^*\, dZ. \tag{4.20}$$

PROOF. Verify that the set H of functions of the form of (4.18), with norm given by (4.19), is a Hilbert space satisfying (4.15) and (4.16).

THEOREM 4E. (General solution of the prediction problem.) *Let $\{X(t), t \in T\}$, be a time series with covariance kernel $K(s,t)$, and let $H(K)$ be the corresponding reproducing kernel Hilbert space. Between $L_2(X(t), t \in T)$ and $H(K)$ there exists a one-one inner product preserving linear mapping under which $X(t)$ and $K(\cdot,t)$ are mapped into one another. Denote by $(h, X)_K$ the random variable in $L_2(X(t), t \in T)$ which corresponds under the mapping to the function $h(\cdot)$ in $H(K)$. Then the general solution to the prediction problem may be written as follows. If Z is a random variable with finite second moment, and if*

$$\rho_Z(t) = E[ZX(t)], \tag{4.21}$$

then

$$E^*[Z \mid X(t), t \varepsilon T] = (\rho_Z, X)_K \tag{4.22}$$

with mean square error of prediction given by

$$E[\mid Z - E^*[Z \mid X(t), t \varepsilon T] \mid^2] = E \mid Z \mid^2 - (\rho_Z, \rho_Z)_K . \tag{4.23}$$

Theorem 4E represents a coordinate free solution of the prediction problem. The usual methods of explicitly writing optimum predictors, using either eigenfunction expansions, Green's functions (impulse response function), or (power) spectral density functions, are merely methods of writing down the reproducing kernel inner product corresponding to the covariance kernel $K(s, t)$ of the observed time series.

The validity of Theorem 4E follows immediately from the definition of the concepts involved. However, it may be instructive to give a proof of the theorem, using the following properties of the mapping $(h, X)_K$. For any functions g and h in $H(K)$ and random variables Z with finite second moment it holds that

$$E[(h, X)_K(g, X)_K] = (h, g)_K \tag{4.24}$$

$$E[Z(h, X)_K] = (\rho_Z, h)_K , \tag{4.25}$$

in which $\rho_Z(t) = E[ZX(t)]$. Now a random variable in $L_2(X(t), t \varepsilon T)$ may be written $(h, X)_K$ for some h in $H(K)$. Consequently the mean square error between any linear functional $(h, X)_K$ and Z may be written

$$\begin{aligned} E[\mid (h, X)_K - Z \mid^2] &= E[(h, X)_K^2] + E[Z^2] - 2E[Z(h, X)_K] \\ &= E[Z^2] + (h, h)_K - 2(\rho_Z, h)_K \\ &= E[Z^2] - (\rho_Z, \rho_Z)_K + (h - \rho_Z, h - \rho_Z)_K . \end{aligned} \tag{4.26}$$

From (4.26) it is immediate that $(\rho_Z, X)_K$ is the minimum mean square error linear predictor of Z, with mean square prediction error equal to $E[Z^2] - (\rho_Z, \rho_Z)_K$. The proof of Theorem 4E is complete.

5. Examples of reproducing kernel Hilbert space representations. In this section we give the reproducing kernel Hilbert space representation of a time series $\{X(t), t \varepsilon T\}$ under a variety of standard assumptions.

EXAMPLE 5A. Suppose $T = \{1, 2, \cdots, N\}$ for some positive integer N, and that the covariance kernel K is given by a symmetric positive definite matrix $\{K_{ij}\}$ with inverse $\{K^{ij}\}$. The corresponding reproducing kernel space $H(K)$ consists of all N-dimensional vectors $f = (f_1, \cdots, f_N)$ with inner product

$$(f, g)_K = \sum_{s,t=1}^{N} f_s K^{st} g_t . \tag{5.1}$$

To prove (5.1) one need only verify that the reproducing property holds: for $u = 1, \cdots, N$,

$$(f, K_{\cdot u})_K = \sum_{s,t=1}^{N} f_s K^{st} K_{tu} = \sum_{s=1}^{N} f_s \delta(s, u) = f_u .$$

The inner product may also be written as a ratio of determinants:

$$(5.2) \qquad (f, g)_K = -\begin{vmatrix} K_{11} & \cdots & K_{1N} & f_1 \\ \vdots & \cdots & \vdots & \vdots \\ K_{N1} & & K_{NN} & f_N \\ g_1 & & g_N & 0 \end{vmatrix} \div \begin{vmatrix} K_{11} & \cdots & K_{1N} \\ \vdots & & \\ K_{N1} & \cdots & K_{NN} \end{vmatrix}.$$

To prove (5.2) one again need only verify the reproducing property. In the case in which the covariance matrix K is singular, one may define the corresponding reproducing kernel inner product in terms of the *pseudo-inverse* of the matrix K (see Greville [15] for a discussion of the notion of pseudo-inverse).

Although (5.1) provides a formula for the reproducing kernel inner product in terms of the inverse of the covariance matrix, it is to be emphasized that one need not necessarily invert the covariance matrix in order to find the reproducing kernel inner product. The point of introducing the reproducing kernel inner product is that the inversion of the covariance matrix is usually an intractable problem, and one should look instead to evaluate that for which one would use the inverse K^{-1}; namely, the evaluation of inner products in the reproducing kernel space. Various iterative methods of evaluating these inner products can be given (see [39] or [40]). This observation is undoubtedly not as important in the case of discrete parameter time series as it is in the case of multiple time series and continuous parameter time series.

EXAMPLE 5B. *Autoregressive schemes (discrete parameter).* A discrete parameter weakly stationary time series $X(t)$ is said to satisfy an autoregressive scheme of order m if $X(t)$ is the solution of the stochastic difference equation

$$(5.3) \qquad L_t X(t) = \sum_{k=0}^{m} a_k X(t - k) = \eta(t)$$

where $a_0, \cdots, a_m$ are given constants, and $\{\eta(t)\}$ is an orthonormal sequence of random variables. We now show that given observations $\{X(t), t = 1, 2, \cdots, N\}$ the reproducing kernel Hilbert space $H(K)$ corresponding to the covariance kernel K of the observations consists of all N-vectors $f = ((f(1), \cdots, f(N))$ with inner product given by

$$(5.4) \qquad (f, g)_K = \sum_{t=m+1}^{N} \{L_t f(t)\}\{L_t g(t)\} + \sum_{j,k=1}^{m} d_{jk} f(j) g(k)$$

where the matrix $D = \{d_{jk}\}$ has an inverse $D^{-1} = \{d^{jk}\}$ with general term

$$(5.5) \qquad d^{jk} = K(j - k) = E[X(j)X(k)].$$

In the case that $N \geqq 2m$, an explicit expression for d_{jk} is given by

$$(5.6) \qquad d_{jk} = \sum_{u=1}^{\min(j,k)} \{a_{j-u} a_{k-u} - a_{u+m-j} a_{u+m-k}\}.$$

In particular for a first order autoregressive scheme and $N \geqq 2$

$$(f, g)_K = (a_0^2 - a_1^2)f(1)g(1) + \sum_{t=2}^{N} \{a_0 f(t) + a_1 f(t-1)\}\{a_0 g(t) + a_1 g(t-1)\}. \tag{5.7}$$

For a second order autoregressive scheme and $N \geqq 4$

$$\begin{aligned}(f, g)_K = {} & (a_0^2 - a_2^2)\{f(1)g(1) + f(2)g(2)\} \\ & + (a_0 a_1 - a_1 a_2)\{f(1)g(2) + f(2)g(1)\} \\ & + \sum_{t=3}^{N} \{a_0 f(t) + a_1 f(t-1) + a_2 f(t-2)\} \\ & \cdot \{a_0 g(t) + a_1 g(t-1) + a_2 g(t-2)\}.\end{aligned} \tag{5.8}$$

One can give a purely algebraic proof of (5.4). However a simpler proof can be given if one uses certain facts from probability theory. Let us suppose that $X(1), \cdots, X(N)$ are jointly normally distributed random variables with covariance matrix $K_N = \{K_{s,t}\}$ with inverse matrix $K_N^{-1} = \{K^{st}\}$. Then the joint probability density function of $X(1), \cdots, X(N)$ may be written

$$f_{X(1),\cdots,X(N)}(x_1, \cdots x_n) = \{(2\pi)^N \mid K_N \mid\}^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x, x)_{K_N}\} \tag{5.9}$$

where $|K_N|$ is the determinant of K_N, and the inner product $(x, x)_{K_N}$ is defined by the right hand side of (5.1). On the other hand, if $X(1), \cdots, X(N)$ satisfy the difference equation $L_t X(t) = \eta(t)$, where $\eta(1), \cdots, \eta(N)$ are independent normal random variables with means 0 and variance 1, then

$$\begin{aligned}& f_{X(1),\cdots,X(m),\eta(m+1),\cdots,\eta(N)}(x_1, \cdots, x_m, y_{m+1}, \cdots, y_N) \\ & \qquad = \{(2\pi)^N |K_m|\}^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\{(x, x)_{K_m} + \sum_{j=m+1}^{N} y_j^2\}\right].\end{aligned} \tag{5.10}$$

Transforming from

$$(X(1), \cdots, X(m), \eta(m+1), \cdots, \eta(N)) \text{ to } (X(1), \cdots, X(N))$$

by the linear transformation $L_t X(t) = \eta(t)$, $t = m+1, \cdots, N$, it follows from (5.10) that

$$\begin{aligned}& f_{X(1),\cdots,X(N)}(x_1, \cdots, x_N) = \{(2\pi)^N |K_m|\}^{-\frac{1}{2}} a_0^{N-m} \\ & \qquad \cdot \exp\left[-\frac{1}{2}\left\{(x, x)_{K_m} + \sum_{j=m+1}^{N} |L_t x_t|^2\right\}\right].\end{aligned} \tag{5.11}$$

Comparing (5.9) and (5.11) it follows that for any N-vector x

$$(x, x)_{K_N} = (x, x)_{K_m} + \sum_{j=m+1}^{N} |L_t x_t|^2 \tag{5.12}$$

which is equivalent to (5.4).

To prove (5.6), define the function $e_j(t)$ by $e_j(t) = 1$ or 0 according as $t = j$ or $t \neq j$. Since the time series $X(t)$ is stationary,

$$(e_j , e_k)_K = (e_{N-j+1} , e_{N-k+1})_K . \tag{5.13}$$

For $1 \leqq j, k \leqq m$, defining $a_j = 0$ for $j < 0$ or $j > m$,

$$\begin{aligned}(e_{N-j+1} , e_{N-k+1})_K &= \sum_{t=m+1}^{N} L_t(e_{N-j+1})L_t(e_{N-k+1}) \\ &= \sum_{t=m+1}^{N} a_{j+t-N-1}a_{k+t-N-1} \\ &= \sum_{u=1}^{N-m} a_{j-u}a_{k-u}\end{aligned} \tag{5.14}$$

while

$$\begin{aligned}(e_j , e_k)_K &= d_{jk} + \sum_{t=m+1}^{N} L_t(e_j)L_t(e_k) \\ &= d_{jk} + \sum_{t=m+1}^{N} a_{t-j}a_{t-k} \\ &= d_{jk} + \sum_{u=1}^{N-m} a_{u+m-j}a_{u+m-k}.\end{aligned} \tag{5.15}$$

From (5.13), (5.14), and (5.15), we obtain (5.6).

From (5.4) and (5.6) one may obtain the inverse matrix of the covariance matrix of an autoregressive scheme (see Siddiqui [49] and references cited there).

EXAMPLE 5C: *Autoregressive schemes (continuous parameter)*. We next consider the reproducing kernel Hilbert space corresponding to the covariance kernel of an autoregressive scheme $X(t)$ observed over a finite interval $a \leqq t \leqq b$.

A continuous parameter stationary time series $X(t)$ is said to be an autoregressive scheme of order m if its covariance function $R(u) = E[X(t)X(t + u)]$ may be written (see Doob [D1], p. 542)

$$R(s - t) = \int_{-\infty}^{\infty} \frac{e^{i(s-t)\omega}}{2\pi \left| \sum_{k=0}^{m} a_k (i\omega)^{m-k} \right|^2} d\omega \tag{5.16}$$

where the polynomial $\sum_{k=0}^{m} a_k z^{m-k}$ has no zeros in the right half of the complex z-plane. It may be shown that given observations of such a time series over a finite interval $a \leqq t \leqq b$, the corresponding reproducing kernel Hilbert space contains all functions $h(t)$ on $a \leqq t \leqq b$ which are continuously differentiable of order m. The reproducing kernel inner product is given by

$$(h, g)_K = \int_a^b (L_t h)(L_t g)\, dt + \sum_{j,k=0}^{m-1} d_{j,k} h^{(j)}(a) g^{(k)}(a) \tag{5.17}$$

where

$$L_t h = \sum_{k=0}^{m} a_k h^{(m-k)}(t) \tag{5.18}$$

$$\{d_{j,k}\}^{-1} = \left\{ \frac{\partial^{j+k}}{\partial t^j \partial u^k} R(t-u) \Big|_{t=a, u=a} \right\}. \tag{5.19}$$

The first and second autoregressive schemes are of particular importance.

A stationary time series $X(t)$ is said to satisfy a first order autoregressive scheme if it is the solution of a first order linear differential equation whose input is white noise $\eta'(t)$ (the symbolic derivative of a process $\eta(t)$ with independent stationary increments):

$$(dX/dt) + \beta X = \eta'(t). \tag{5.20}$$

It should be remarked that from a mathematical point of view (5.20) should be written

$$dX(t) + \beta X(t)dt = d\eta(t). \tag{5.21}$$

Even then, by saying that $X(t)$ satisfies (5.20) or (5.21) we mean that

$$X(t) = \int_{-\infty}^{t} H(t-s)\, d\eta(s) \tag{5.22}$$

where $H(t-s) = e^{-\beta(t-s)}$ is the one-sided Green's function of the differential operator $L_t f = f'(t) + \beta f(t)$.

The covariance function of the stationary time series $X(t)$ is

$$R(t-u) = (1/2\beta)e^{-\beta|u-t|} \tag{5.23}$$

The corresponding reproducing kernel Hilbert space $H(K)$ contains all differentiable functions. The inner product is given by

$$(h, g)_k = \int_a^b (h' + \beta h)(g' + \beta g)\, dt + 2\beta h(a)g(a). \tag{5.24}$$

More generally, corresponding to the covariance function

$$K(s, t) = Ce^{-\beta|s-t|} \tag{5.25}$$

the reproducing kernel inner product is

$$\begin{aligned}(h, g)_K &= \frac{1}{2\beta C}\left\{\int_a^b (h' + \beta h)(g' + \beta g)\, dt + 2\beta h(a)g(a)\right\} \\ &= \frac{1}{2\beta C}\int_a^b (h'g' + \beta^2 hg)\, dt + \frac{1}{2C}\{h(a)g(a) + h(b)g(b)\}.\end{aligned} \tag{5.26}$$

The random variable $(h, X)_K$ in $L_2(X(t), a \leqq t \leqq b)$ corresponding to $h(\cdot)$ in

$H(K)$ may be written

$$(h, X)_K = \frac{1}{2\beta C}\left\{\beta^2 \int_a^b h(t)X(t)\,dt + \int_a^b h'(t)\,dX(t)\right\} + \frac{1}{2C}\{h(a)X(a) + h(b)X(b)\}. \tag{5.27}$$

Note that $X'(t)$ does not exist in any rigorous sense; consequently we write $dX(t)$ where $X'(t)\,dt$ seems to be called for. It can be shown that (5.27) makes sense. In the case that $h(\cdot)$ is twice differentiable, one may integrate by parts and write

$$\int_a^b h'(t)\,dX(t) = h'(b)X(b) - h'(a)X(a) - \int_a^b X(t)h''(t)\,dt. \tag{5.28}$$

A stationary time series $X(t)$ is said to satisfy a second order autoregressive scheme if it is the solution of a second order linear differential equation whose input is white noise $\eta'(t)$:

$$(d^2X/dt^2) + 2\alpha(dX/dt) + \gamma^2 X = \eta'(t). \tag{5.29}$$

If $\omega^2 = \gamma^2 - \alpha^2 > 0$, the covariance function of the time series is

$$R(t - u) = \frac{e^{-\alpha|u-t|}}{4\alpha\gamma^2}\left\{\cos\omega(u - t) + \frac{\alpha}{\omega}\sin\omega|u - t|\right\}. \tag{5.30}$$

The corresponding reproducing kernel Hilbert space contains all twice differentiable functions on the interval $a \leqq t \leqq b$ with inner product

$$(h, g)_K = \int_a^b (h'' + 2\alpha h' + \gamma^2 h)(g'' + 2\alpha g' + \gamma^2 g)\,dt + 4\alpha\gamma^2 h(a)g(a) + 4\alpha h'(a)g'(a). \tag{5.31}$$

To write an expression for $(h, X)_K$, one uses the same considerations as in (5.27).

Other examples of reproducing kernel Hilbert spaces are given in [39] and [40].

6. Regression analysis of time series with known covariance function. The theory of regression analysis (and of the general linear hypothesis) plays a central role in statistical theory. In this section we show how to solve certain standard problems of regression analysis in cases in which the observations possess properties of dependence or continuity. For a discussion of the history and literature of regression analysis the reader is referred to Wold [58].

The classical problem of regression analysis may be posed as follows. Given (i) observations $X(t)$, $t = 1, \cdots, N$, with known covariance kernel

$$K(s, t) = \text{Cov}\,[X(s), X(t)] \tag{6.1}$$

and mean value function $m(t) = E[X(t)]$ of the form

$$m(t) = \beta_1 w_1(t) + \cdots + \beta_q w_q(t) \tag{6.2}$$

where $w_1(\cdot), \cdots, w_q(\cdot)$ are known functions, and $\beta_1, \cdots, \beta_q$ are unknown real

numbers, and (ii) a linear function

$$\psi(\beta) = \psi_1\beta_1 + \cdots + \psi_q\beta_q \tag{6.3}$$

of the parameters, where $\psi_1, \cdots, \psi_q$ are known constants. Estimate $\psi(\cdot)$ by an estimate which (i) is *linear* in the observations in the sense that it is of the form $\sum_{t=1}^{N} c_t X(t)$ for some real numbers $c_1, \cdots, c_N$, (ii) is an *unbiased* estimate of $\psi(\cdot)$ in the sense, that for all $\beta = (\beta_1, \cdots, \beta_q)$,

$$E_\beta\left[\sum_{t=1}^{N} c_t X(t)\right] = \sum_{t=1}^{N} c_t m(t) = \sum_{j=1}^{q} \beta_j \sum_{t=1}^{N} c_t w_j(t) = \psi(\beta), \tag{6.4}$$

and (iii) has variance

$$\operatorname{Var}\left[\sum_{t=1}^{N} c_t X(t)\right] = \sum_{s,t=1}^{N} c_s K(s, t) c_t \tag{6.5}$$

equal to the minimum variance of any unbiased linear estimate.

The problem of finding the minimum variance unbiased linear estimate of a linear parametric function $\psi(\beta)$ can be posed as a problem involving the minimization of a *quadratic form* subject to linear restraints. Define $K = \{K(s, t)\}$,

$$W = \begin{bmatrix} w_1(1) & \cdots & w_q(1) \\ \vdots & \cdots & \vdots \\ w_1(N) & \cdots & w_q(N) \end{bmatrix}, \quad \psi = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_q \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \tag{6.6}$$

and let c' denote the transpose of a (column) vector c. The unbiasedness condition (6.4) can be stated in matrix form as

$$c'W = \psi'. \tag{6.7}$$

The problem of finding the minimum variance unbiased linear estimate can now be posed as follows: find the vector c which minimizes the quadratic form $c'Kc$, subject to the constraints $c'W = \psi'$ (compare Bush and Olkin [7]).

THEOREM 6A. *Let K be a positive definite $n \times n$ symmetric matrix, W be an $n \times q$ matrix, and ψ a q-vector. Assume that*

$$V = W'K^{-1}W \tag{6.8}$$

is non-singular. The n-vector c^* which minimizes the quadratic form $c'Kc$ among all n-vectors c satisfying $W'c = \psi$ is given by

$$c^* = K^{-1}WV^{-1}\psi \tag{6.9}$$

and the minimum value of the quadratic form is given by

$$c^{*\prime}Kc^* = \psi'V^{-1}\psi. \tag{6.10}$$

PROOF. One easily verifies that the vector c^* defined by (6.9) satisfies the restraint $W'c = \psi$, and that (6.10) holds. To complete the proof we show that for any n-vector c such that $c'W = \psi'$ it holds that

$$c'Kc \geqq \psi'V^{-1}\psi. \tag{6.11}$$

Now for any q-vector z, letting $y = Wz$,

$$c'Kc \geqq \frac{(c'y)^2}{y'K^{-1}y} = \frac{(c'Wz)^2}{z'Vz} = \frac{(\psi'z)^2}{z'Vz}. \tag{6.12}$$

Taking the supremum of the right side of (6.12) over all q-vectors z, one obtains (6.11), since

$$\sup_z [(\psi'z)^2/z'Vz] = \psi'V^{-1}\psi. \tag{6.13}$$

From Theorem 6A, one immediately obtains Theorem 6B.

THEOREM 6B. *The minimum variance linear unbiased estimate of a parametric function $\psi(\beta)$ is*

$$\psi^* = c^{*\prime}X = \psi'V^{-1}(W'K^{-1}X). \tag{6.14}$$

The variance of ψ^* is given by

$$\text{Var}\,[\psi^*] = c^{*\prime}Kc^* = \psi'V^{-1}\psi. \tag{6.15}$$

In particular, the vector $\beta^{*\prime} = (\beta_1^*, \cdots, \beta_q^*)$ of minimum variance unbiased linear estimates of $\beta_1, \cdots, \beta_q$ may be written

$$\beta^* = V^{-1}(W'K^{-1}X) \tag{6.16}$$

with covariance matrix

$$\{\text{Cov}\,[\beta_i^*, \beta_j^*]\} = V^{-1}. \tag{6.17}$$

The foregoing treatment of the problem of regression analysis with known covariance function depended very much on the assumptions that there were only a finite number of observations, and that the matrices K and V were non-singular. We now show how to relax these assumptions by using the reproducing kernel Hilbert space representation of a time series. The results we now state include as special cases the results which were first obtained by Grenander ([12], [14]).

Let $\{X(t), t \varepsilon T\}$ be a time series whose covariance kernel $K(s, t) = \text{Cov}\,[X(s), X(t)]$ is known and whose mean value function $m(t) = E[X(t)]$ is only assumed to belong to a known class M. Let $H(K)$ be the reproducing kernel Hilbert space corresponding to K. Assume that M is a subset of $H(K)$. It may be shown that between $L_2(X(t), t \varepsilon T)$ and $H(K)$ there exists a one-one linear mapping with the following properties: if $(h, X)_K$ denotes the random variable in $L_2(X(t), t \varepsilon T)$ which corresponds under the mapping to the function h in $H(K)$, then for every t in T, and h and g in $H(K)$,

$$(K(\cdot, t), X)_K = X(t), \tag{6.18}$$

$$E_m[(h, X)_K] = (h, m)_K, \qquad \text{for all } m \text{ in } M, \tag{6.19}$$

$$\text{Cov}\,[(h, X)_K, (g, X)_K] = (h, g)_K. \tag{6.20}$$

The subscript m on an expectation operator is written to indicate that the expectation is computed under the assumption that $m(\cdot)$ is the true mean value function.

If T is finite, and K is non-singular, then $(h, X)_K = h'K^{-1}X$. For other examples of $(h, X)_K$, see Section 5.

We are interested in estimating various functionals $\psi(m)$ of the true mean value function $m(\cdot)$ by estimates which (i) are linear in the observations $\{X(t), t \varepsilon T\}$ in the sense that they belong to $L_2(X(t), t \varepsilon T)$, (ii) are *unbiased* and (iii) have *minimum variance* among all linear unbiased estimates. A functional $\psi(m)$ is said to be linearly estimable if it possesses an unbiased linear estimate $(g, X)_K$. Since

$$E_m[(g, X)_K] = (g, m)_K = \psi(m), \qquad \text{for all } m \text{ in } M \tag{6.21}$$

it follows that $\psi(m)$ is linearly estimable if and only if there exists a function g in $H(K)$ satisfying (6.21). Now the variance of a linear estimate is given by

$$\text{Var}\,[(g, X)_K] = (g, g)_K . \tag{6.22}$$

Consequently finding the minimum variance unbiased linear estimate $\psi^* = (g^*, X)_K$ of $\psi(m)$ is equivalent to finding that function g^* in $H(K)$ which has minimum norm among all functions g satisfying the restraint (6.21). To find the vector g^* with minimum norm it suffices to find any vector g satisfying (6.21). Then the projection

$$g^* = E^*[g \mid \bar{M}], \tag{6.23}$$

of g onto the smallest Hilbert subspace $\bar{M}$ containing M, satisfies (6.21) and has mnimum norm among all vectors satisfying (6.21).

THEOREM 6C. *The uniformly minimum variance unbiased linear estimate ψ^* of a linearly estimable function $\psi(m)$ is given by*

$$\psi^* = (E^*[g \mid \bar{M}], X)_K \tag{6.24}$$

with variance

$$\text{Var}\,[\psi^*] = \|E^*[g \mid \bar{M}]\|_K^2 , \tag{6.25}$$

where g is any function satisfying (6.21), *$\bar{M}$ is the smallest Hilbert subspace of $H(K)$ containing M, and $E^*[g \mid \bar{M}]$ denotes the projection onto $\bar{M}$ of g. In particular, the uniformly minimum variance unbiased linear estimate $m^*(t)$ of the value $m(t)$ at a particular point t of the mean value function $m(\cdot)$ is given by*

$$m^*(t) = (E^*[K(\cdot, t) \mid \bar{M}], X)_K \tag{6.26}$$

since

$$m(t) = (K(\cdot, t), m)_K . \tag{6.27}$$

In the special case that M consists of all functions $m(t)$ of the form of (6.2), and the matrix

$$V = \begin{bmatrix} (w_1, w_1)_K & \cdots & (w_1, w_q)_K \\ \vdots & & \\ (w_q, w_1)_K & \cdots & (w_q, w_q)_K \end{bmatrix} \tag{6.28}$$

is non-singular, then

$$\beta^* = V^{-1}\begin{bmatrix}(w_1, X)_K \\ \vdots \\ (w_q, X)_K\end{bmatrix}. \tag{6.29}$$

One may write an explicit formula for the minimum variance unbiased linear estimate ψ^* of a linear parametric function $\psi(\beta) = \psi_1\beta_1 + \cdots + \psi_q\beta_q$ as follows, where $V_{ij} = (w_i, w_j)_K$;

$$\psi^* = -\begin{vmatrix} V_{11} & \cdots & V_{1q} & (X, w_1)_K \\ \vdots & \cdots & \vdots & \vdots \\ V_{q1} & \cdots & V_{qq} & (X, w_q)_K \\ \psi_1 & \cdots & \psi_q & 0 \end{vmatrix} \div \begin{vmatrix} V_{11} & \cdots & V_{1q} \\ \vdots & \cdots & \vdots \\ V_{q1} & \cdots & V_{qq} \end{vmatrix}. \tag{6.30}$$

It should be noted that the proof of Theorem 6C is exactly the same in spirit as the proof of the Gauss-Markov theorem given in Scheffé ([46], p. 14). The point of Theorem 6C is that it enables one to develop a theory of regression analysis and analysis of variance for cases in which one has an infinite number of observations. In particular, we state the analogues of certain basic results on simultaneous confidence intervals (Scheffé [46], p. 68) and hypothesis testing (Scheffé [46], p. 31).

Hypothesis testing and simultaneous confidence bands for mean value functions. If the time series $X(t)$ is assumed to be normal, or if all linear functionals $(h, X)_K$ may be assumed to be approximately normally distributed, then one may state a confidence band for the entire mean value function $m(\cdot)$ as follows. Given a confidence level α, let $C_q(\alpha)$ denote the α percentile of the χ^2 distribution with q degrees of freedom; in symbols, $P[\chi_q^2 \geqq C_q(\alpha)] = \alpha$.

We now show that if the smallest space $\bar{M}$ containing all mean value functions has finite dimension q, then

$$m^*(t) - [C_q(\alpha)]^{\frac{1}{2}}\sigma[m^*(t)] \leqq m(t) \leqq m^*(t) + [C_q(\alpha)]^{\frac{1}{2}}\sigma[m^*(t)] \quad \text{for all } t \text{ in } -\infty < t < \infty \tag{6.31}$$

is a simultaneous confidence band for all values of the mean value function with a level of significance not less that α; that is, if $m(\cdot)$ is the true mean value function then (6.31) holds with a probability greater than or equal to α.

To prove (6.31) we prove more generally the following theorem.

THEOREM 6C. (Simultaneous confidence interval of significance level α for all estimable functions (m, g).) *If $\bar{M}$ has dimension q then for all m in $\bar{M}$*

$$P_m\left[\sup_{g \in H(K)} \frac{|(X, E^*[g \mid \bar{M}])_K - (m, g)_K|^2}{\operatorname{Var}[(X, E^*[g \mid \bar{M}])_K]} \leqq C_q(\alpha)\right] = \alpha. \tag{6.32}$$

PROOF. Let $w_1, \cdots, w_q$ be orthonormal functions which span $\bar{M}$. Then we may write $m = \beta_1 w_1 + \cdots + \beta_q w_q$ where $\beta_j = (m, w_j)$ is a function of m. Further, $(m, g)_K = \alpha_1\beta_1 + \cdots + \alpha_q\beta_q$, $(X, E^*[g \mid \bar{M}])_K = \alpha_1\beta_1^* + \cdots + \alpha_q\beta_q^*$,

Var $[(X, E^*[g \mid \bar{M}])_K] = \sum_{j=1}^q \alpha_j^2$, where $\alpha_j = (w_j, g)_K$ and $\beta_j^* = (X, w_j)_K$. Next the random variable appearing in (6.32) is equal to

$$\sup_{-\infty<\alpha_1,\cdots,\alpha_q<\infty} \frac{\left|\sum_{j=1}^q \alpha_j(\beta_j^* - \beta_j)\right|^2}{\sum_{j=1}^q \alpha_j^2} = \sum_{j=1}^q (\beta_j^* - \beta_j)^2 \tag{6.33}$$

which is distributed as χ_q^2 (compare Scheffé [46], p. 416).

Similarly one may prove the following theorem.

THEOREM 6D. *Given a q-dimensional subspace M of $H(K)$, and a q'-dimensional subspace M' of M, to test the composite null hypothesis $H_0 : m(\cdot) \varepsilon M'$, against the composite alternative hypothesis, $H_1 : m(\cdot) \varepsilon M$, one may use the statistic*

$$\Delta = \|m_M^*(t) - m_{M'}^*(t)\|_K^2 \tag{6.34}$$

where $m_M^*(t)$ $[m_{M'}^*(t)]$ denotes the minimum variance unbiased linear estimate of $m(t)$ under the hypothesis $H_1[H_0]$. Under H_0, Δ is distributed as χ^2 with $q - q'$ degrees of freedom.

In the special case that M consists of all functions $m(t)$ of the form (6.2), and M' consists of all functions in M for which $\beta_j = 0$ for $j = q' + 1, \cdots, q$, then the statistic Δ may be written

$$\Delta = \sum_{j=q'+1}^q \delta_j \tag{6.35}$$

where, defining $V_{ij} = (w_i, w_j)_K$,

$$\begin{aligned} \delta_j &= \frac{|(w_j - E^*[w_j \mid w_1, \cdots, w_{j-1}], X)_K|^2}{\|w_j - E^*[w_j \mid w_1, \cdots, w_{j-1}]\|_K^2} \\ &= \begin{vmatrix} V_{11} \cdots V_{1,j-1} & (w_1, X)_K \\ \vdots & \\ V_{j1} \cdots V_{j,j-1} & (w_j, X)_K \end{vmatrix}^2 \div \begin{vmatrix} V_{11} \cdots V_{1j} \\ \vdots \quad \cdots \quad \vdots \\ V_{j1} \cdots V_{jj} \end{vmatrix} \begin{vmatrix} V_{11} \quad \cdots \quad V_{1,j-1} \\ V_{j-1,1} \cdots V_{j-1,j-1} \end{vmatrix} \end{aligned} \tag{6.36}$$

The reader may find it illuminating to write out (6.36) in the case that $q = 2$ and $q' = 1$.

Regression analysis when the covariance function of the observations is only known up to a constant factor. Suppose that the covariance function of the time series $\{X(t), t \varepsilon T\}$ is of the form

$$\text{Cov}\,[X(s), X(t)] = \sigma^2 K(s, t)$$

where the kernel $K(s, t)$ is known and σ^2 is an unknown positive constant, and that the mean value function $m(t) = E[X(t)]$ is known to belong to a set M which is a subspace (of dimension q) of $H(K)$, the reproducing kernel Hilbert space corresponding to K. Theorem 6C continues to hold, except that (6.25) should be replaced by

$$\text{Var}_\sigma\,[\psi^*] = \sigma^* \|E^*[g \mid M]\|_K^2. \tag{6.25'}$$

The variance of the estimate ψ^* depends on the unknown parameter σ^2. Therefore one needs to estimate σ^2 in order to know $\mathrm{Var}_\sigma[\psi^*]$. To discuss the estimation of σ^2, we need to distinguish between the case in which the index set T is finite and the case in which T is infinite.

If T is finite, the time series $\{X(t), t \varepsilon T\}$, regarded as a function of t, may be shown to belong to $H(K)$. Further if n is the dimension of $H(K)$, then for all possible mean value functions $m(t)$ and values of σ^2

$$
\begin{aligned}
E[\|X(t) - m(t)\|_K^2] &= n\sigma^2 \\
E[\|m^*(t) - m(t)\|_K^2] &= q\sigma^2 \\
E[\|X(t) - m^*(t)\|_K^2] &= (n - q)\sigma^2.
\end{aligned} \tag{6.37}
$$

Therefore

$$
\sigma^{*2} = (n - q)^{-1} \|X(t) - m^*(t)\|_K^2 \tag{6.38}
$$

is an unbiased estimate of σ^2 (which in the case of normally distributed observations is independent of $m^*(t)$).

If T is infinite, it is possible to estimate σ^2 exactly by forming a sequence of estimates of the form of (6.38) based on a monotone sequence of finite subsets $\{T_n\}$ of T whose limit is dense in T.

7. The probability density functional of a normal process. The prediction and regression problems considered in the foregoing have all involved linear estimates chosen according to a criterion expressed in terms of mean square error. Nevertheless the mathematical tools developed continue to play an important role if one desires to employ other criteria of statistical inference. All modern theories of statistical inference take as their starting point the idea of the probability density function of the observations. Thus in order to apply any principle of statistical inference to problems of time series analysis, it is first necessary to develop the notion of the probability density function (or functional) of a stochastic process. In this section we state a result showing how one may write a formula for the probability density function of a stochastic process which is normal.

Given a normal time series $\{X(t), t \varepsilon T\}$ with known covariance function

$$
K(s, t) = \mathrm{Cov}\,[X(s), X(t)] \tag{7.1}
$$

and mean value function $m(t) = E[X(t)]$, let P_m be the probability measure induced on the space of sample functions of the time series. Next, let m_1 and m_2 be two functions, and let P_1 and P_2 be the probability measures induced by normal time series with the same covariance kernel K, and mean value functions equal to m_1 and m_2 respectively. By the Lebesgue decomposition theorem it follows that there is a set N of P_1-measure 0 and a non-negative P_1-integrable function, denoted by dP_2/dP_1, such that for every measurable set B of sample functions

$$
P_2(B) = \int_B (dP_2/dP_1)\, dP_1 + P_2(BN'). \tag{7.2}
$$

If $P_2(N) = 0$, then P_2 is absolutely continuous with respect to P_1, and dP_2/dP_1 is called the probability density function of P_2 with respect to P_1. Two measures which are absolutely continuous with respect to one another are called *equivalent.* Two measures P_1 and P_2 are said to be *orthogonal* if there is a set N such that $P_1(N) = 0$ and $P_2(N) = 1$.

It has been proved, independently by various authors under various hypotheses (for references, see [40], Section 4), that two normal probability measures are either equivalent or orthogonal. From the point of view of obtaining an explicit formula for the probability density function, the following formulation of this theorem is useful.

THEOREM 7A (Parzen [37], [40]). *Let P_m be the probability measure induced on the space of sample functions of a time series $\{X(t), t \varepsilon T\}$ with covariance kernel K and mean value function m. Assume that either* (i) *T is countable or* (ii) *T is a separable metric space, K is continuous, and the stochastic process $\{X(t), t \varepsilon T\}$ is separable. Let P_0 be the probability measure corresponding to the normal process with covariance kernel K and mean value function $m(t) = 0$. Then P_m and P_0 are equivalent or orthogonal, depending on whether m does or does not belong to the reproducing kernel Hilbert space $H(K)$. If m belongs to $H(K)$, then the probability density functional of P_m with respect to P_0 is given by*

$$f(X, m) = dP_m/dP_0 = \exp\{(X, m)_K - (\tfrac{1}{2})(m, m)_K\}. \tag{7.3}$$

Using the concrete formula for the probability density functional of a normal process provided by (7.3), there is no difficulty in applying the concepts of classical statistical methodology to problems of inference on normal time series. In particular the following theorem may be proved.

THEOREM 7B. *Let $\{X(t), t \varepsilon T\}$ be a normal time series, satisfying the assumptions of Theorem* 7A *with known covariance kernel $K(s, t) = \mathrm{Cov}[X(s), X(t)]$, whose mean value function is only assumed to belong to a known class M. If M is a finite dimensional subspace of the reproducing kernel space $H(K)$, then the maximum likelihood estimate $m^*(\cdot)$, defined as that estimate in the space M of admissible mean value functions such that*

$$f(X, m^*) = \max_{m \varepsilon M} f(X, m), \tag{7.4}$$

exists and is given at each t in T by the right hand side of (6.26).

If M is an infinite dimensional space, then a maximum likelihood estimate does not exist. This is not too surprising, since M is not compact in this case. However, an estimate does exist which is the uniformly minimum variance unbiased linear estimate of the value $m(t)$ at a particular time t of the mean value function; this estimate is given by (6.26).

The theory of reproducing kernel Hilbert spaces turns out to provide a natural tool for treating problems of minimum variance unbiased estimation (see Parzen [37]). Further work along these lines in the case of normal time series is being done by Ylvisaker ([59]).

8. Correlation analysis of regression free stationary time series. In this section, we state some results for discrete parameter time series (a more comprehensive survey is given by Hannan [17]). Many of the results stated may be extended to continuous parameter time series.

We consider a discrete parameter time series $\{X(t), t = 1, 2, \cdots\}$, with zero means, which is weakly stationary of order 4 in the sense that its covariance function

$$R(v) = E[X(t)X(t+v)] \tag{8.1}$$

and its fourth cumulant function

$$\begin{aligned} Q(v_1, v_2, v_3) &= E[X(t)X(t+v_1)X(t+v_2)X(t+v_3)] \\ &\quad - R(v_1)R(v_2 - v_3) - R(v_2)R(v_1 - v_3) - R(v_3)R(v_1 - v_2) \end{aligned} \tag{8.2}$$

are independent of t.

EXAMPLE: *Linear Processes.* A discrete parameter time series $X(t)$ is said to be a linear process, if it may be represented

$$X(t) = \sum_{\alpha=-\infty}^{\infty} w(t-\alpha)\eta(\alpha) \tag{8.3}$$

where $\sum_{\alpha=-\infty}^{\infty} |w(\alpha)| < \infty$, and $\{\eta(\alpha), \alpha = 0, \pm 1, \cdots\}$ is a sequence of independent identically distributed random variables with zero means, finite fourth cumulant λ_4, and second cumulant λ_2. A linear process $X(t)$ is weakly stationary up to order 4, with covariance function, spectral density function, and fourth cumulant function satisfying

$$R(v) = \lambda_2 \sum_{\alpha=-\infty}^{\infty} w(\alpha)w(\alpha+v), \qquad f(\lambda) = \frac{\lambda_2}{2\pi}\left| \sum_{\alpha=-\infty}^{\infty} w(\alpha)e^{-i\lambda\alpha} \right|^2,$$

$$Q(v_1, v_2, v_3) = \lambda_4 \sum_{\alpha=-\infty}^{\infty} w(\alpha)w(\alpha+v_1)w(\alpha+v_2)w(\alpha+v_3), \tag{8.4}$$

$$\sum_{u=-\infty}^{\infty} Q(v_1, u, u+v_2) = \alpha R(v_1)R(v_2), \qquad \alpha = \frac{\lambda_4}{(\lambda_2)^2}.$$

Correlation analysis is concerned with estimating the covariance function $R(v)$, and the normalized covariance (or correlation) function

$$\rho(v) = R(v)/R(0) \tag{8.5}$$

of a stationary time series.

Given observations $\{X(t), t = 1, 2, \cdots, N\}$, one can form the sample covariance function, for $|v| \leqq N - 1$,

$$R_N(v) = \frac{1}{N} \sum_{t=1}^{N-|v|} X(t)X(t+|v|) \tag{8.6}$$

which has mean

$$E[R_N(v)] = [1 - (|v|/N)]R(v). \tag{8.7}$$

As an estimate of $R(v)$, $R_N(v)$ is biased (although asymptotically unbiased). Consequently if we are interested in estimating $R(v)$ it may be preferable to take as our estimate

$$R_N^u(v) = [N/(N - |v|)]R_N(v).$$

Many authors have advocated the use of the unbiased estimate $R_N^u(v)$ in preference to the biased estimate $R_N(v)$. However, it appears to me that $R_N(v)$ is preferable to $R_N^u(v)$ for two reasons: (i) $R_N(v)$ is a positive definite function of v, which is not the case of $R_N^u(v)$; (ii) the mean square error of $R_N(v)$ as an estimate of $R(v)$ is in general less than that of $R_N^u(v)$. That (i) holds is immediate. That (ii) holds is shown in Parzen [42]. It will be seen that for theoretical purposes it is certainly more useful to consider $R_N(v)$ rather than $R_N^u(v)$.

Using the large of large numbers proved in Parzen ([38], pp. 419–420), one may prove the following theorems on consistency of the sample covariance function.

THEOREM 8A. *The sample covariance function of a weakly stationary time series is consistent in quadratic mean, in the sense that, for* $v = 0, 1, \cdots$,

$$\lim_{N\to\infty} E \mid R_N(v) - R(v) \mid^2 = 0 \tag{8.8}$$

if the time series is weakly stationary of order 4, *and satisfies (for* $v = 0, 1, \cdots$)

$$\lim_{N\to\infty} \frac{1}{N} \sum_{s=0}^{N-1} R^2(s) = 0 \tag{8.9}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{s=0}^{N-v-1} Q(v, s, v + s) = 0. \tag{8.10}$$

THEOREM 8B. *The sample covariance function of a weakly stationary time series is strongly consistent, in the sense that, for each* $v = 0, 1, \cdots$,

$$P[\lim_{N\to\infty} R_N(v) = R(v)] = 1 \tag{8.11}$$

if the time series is weakly stationary of order 4 *and satisfies for positive constants* C *and* q

$$\frac{1}{N} \sum_{s=0}^{N-1} R^2(s) \leqq CN^{-q} \qquad \textit{for all } N \tag{8.12}$$

$$\frac{1}{N} \left| \sum_{s=0}^{N-v-1} Q(v, s, v + s) \right| \leqq CN^{-q} \qquad \textit{for all } N. \tag{8.13}$$

In particular, (8.12) and (8.13) hold if it is assumed that

$$\sum_{v=-\infty}^{\infty} |R(v)| < \infty \tag{8.14}$$

$$\sum_{v_1,v_2,v_3,=-\infty}^{\infty} |Q(v_1, v_2, v_3)| < \infty. \tag{8.15}$$

We next obtain expressions for the asymptotic covariance of the sample co-

variance function (for proofs of the following theorem see Bartlett [2], [3] or Parzen [36]).

THEOREM 8C. *Let $X(t)$ be a time series weakly stationary of order 4, with absolutely summable covariance and fourth cumulant functions (that is, (8.14) and (8.15) hold). Then the sample covariance function $R_N(v)$ has asymptotic covariance, for any non-negative integers v_1 and v_2,*

$$\lim_{N\to\infty} N \operatorname{Cov}[R_N(v_1), R_N(v_2)] = D(v_1, v_2) \tag{8.16}$$

where we define

$$D(v_1, v_2) = \sum_{u=-\infty}^{\infty} \{R(u)R(u + v_2 - v_1) + R(u)R(u + v_2 + v_1) + Q(v_1, u, u + v_2)\}. \tag{8.17}$$

For a linear process with spectral density function $f(\cdot)$

$$D(v_1, v_2) = 4\pi \int_{-\pi}^{\pi} \cos \lambda v_1 \cos \lambda v_2 f^2(\lambda) d\lambda + \alpha \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cos \lambda v_1 \cos \lambda v_2 f(\lambda_1) f(\lambda_2)\, d\lambda_1\, d\lambda_2 . \tag{8.18}$$

In particular, the variance of $R_N(v)$ is approximately given by

$$\operatorname{Var}[R_N(v)] = \frac{4\pi}{N} \int_{-\pi}^{\pi} \cos^2 \lambda v\, f^2(\lambda)\, d\lambda + \frac{\alpha}{N} R^2(v) \geqq \frac{2 + \alpha}{N} R^2(v). \tag{8.19}$$

The mean square error of $R_N(v)$ as an estimate of $R(v)$ is given by

$$E\,|R_N(v) - R(v)|^2 = \operatorname{Var}[R_N(v)] + \left(\frac{v}{N}\right)^2 R^2(v). \tag{8.20}$$

It was empirically observed by M. G. Kendall that the sample covariance function (traditionally called the *observed correlogram*) fails to damp down to 0 for increasing values of v, although the true covariance function $R(v)$ does damp down to 0 as v tends to ∞. This fact is borne out theoretically by (8.19) and (8.20), which show that the coefficient of mean square error $E\,|R_N(v) - R(v)|^2 / R^2(v)$ is of the order of $1/N$ for all lags v of the sample covariance function.

One may state in a variety of ways conditions under which *the sample covariance function $R_N(v)$ is asymptotically normal* in the sense that for every choice of lags $v_1, \cdots, v_k$ and real numbers $u_1, \cdots, u_k$,

$$E[\exp i\{u_1 N^{\frac{1}{2}}(R_N(v_1) - E[R_N(v_1)]) + \cdots + u_k N^{\frac{1}{2}}(R_N(v_k) - E[R_N(v_k)])\}] \to \exp\left[-\frac{1}{2}\left\{\sum_{i,j=1}^{k} u_i D(v_i, v_j)\, u_i\right\}\right] \tag{8.21}$$

as $N \to \infty$ (see Walker [54], Lomnicki and Zaremba [29], Parzen [35]). In particular, (8.21) holds if $X(t)$ is a linear process.

As an estimate of the correlation function $\rho(v)$ we take the sample correlation function

$$\rho_N(v) = R_N(v)/R_N(0). \tag{8.22}$$

We do not discuss here the question of the best definition of the correlation function for short series. For a discussion of this problem, and references to the literature, see Weinstein [55].

By standard large sample statistical theory one readily obtains, from (8.21) and (8.16), the following theorem.

THEOREM 8D. *If $X(t)$ is a linear process, then $\rho_N(v)$ is asymptotically normal with asymptotic covariances satisfying, as $N \to \infty$,*

$$NE[\rho_n(v_1) - \rho(v_1), \rho_N(v_2) - \rho(v_2)] \to d(v_1, v_2) \tag{8.23}$$

where we define

$$d(v_1, v_2) = 4\pi \int_{-\pi}^{\pi} d\lambda \bar{f}^2(\lambda)\{\cos \lambda v_1 - \rho(v_1)\}\{\cos \lambda v_2 - \rho(v_2)\}, \tag{8.24}$$

$$\bar{f}(\lambda) = \frac{f(\lambda)}{R(0)}. \tag{8.25}$$

REMARK. It should be noted that while the variance of the sample covariance function $R_N(v)$ of a linear process depends on α, the variance of the sample correlation function $\rho_N(v)$ does not.

PROOF. Using only the first few terms of the Taylor series expansion one obtains that

$$\frac{x}{y} - \frac{x_0}{y_0} = \frac{y}{y_0^3}\{(y - y_0)x_0 - (x - x_0)y_0\} + 0(|x - x_0|^2 + |y - y_0|^2) \tag{8.26}$$

Consequently, if X_n, Y_n, and Z_n are sequences of random variables, and x_0, y_0, and z_0 are constants, such that

$$n^{\frac{1}{2}}(X_n - x_0), \qquad n^{\frac{1}{2}}(Y_n - y_0), \qquad n^{\frac{1}{2}}(Z_n - z_0)$$

are jointly asymptotically normal it follows that

$$n^{\frac{1}{2}}\left(\frac{X_n}{Y_n} - \frac{x_0}{y_0}\right), \qquad n^{\frac{1}{2}}\left(\frac{Z_n}{Y_n} - \frac{z_0}{y_0}\right)$$

are jointly asymptotically normal with asymptotic covariance satisfying

$$\begin{aligned} y_0^4 n \operatorname{Cov}\left[\frac{X_n}{Y_n} - \frac{x_0}{y_0}, \frac{Z_n}{Y_n} - \frac{z_0}{y_0}\right] &\to x_0 z_0 E[(Y_n - y_0)^2] \\ + y_0^2 E[(X_n - x_0)(Z_n - z_0)] - x_0 y_0 E[(Y_n - y_0)(Z_n - z_0)] \\ - z_0 y_0 E[(Y_n - y_0)(X_n - x_0)]. \end{aligned} \tag{8.27}$$

Applying these results to the present case it follows that the sample correlations

$\rho_N(v)$ are asymptotically normal with asymptotic covariances satisfying

$$R^4(0)N \text{ Cov } [\rho_N(v_1) - \rho(v_1), \rho_N(v_2) - \rho(v_2)]$$

$$= R(v_1)R(v_2)D(0,0) + R^2(0)D(v_1, v_2) - R(0)R(v_1)D(0, v_2) - R(0)R(v_2)D(0, v_1). \tag{8.28}$$

From (8.28) and (8.18), one obtains (8.24).

We are now in a position to obtain confidence intervals for, or test hypotheses about, a correlation coefficient $\rho(v)$. From Theorem 8D it follows that the sample correlation coefficient $\rho_N(v)$ may be regarded as being normally distributed with mean $\rho(v)$ and variance equal to $d(v)/N$ where we define

$$d(v) = 4\pi \int_{-\pi}^{\pi} d\lambda \bar{f}^2(\lambda)\{\cos \lambda v - \rho(v)\}^2. \tag{8.29}$$

Now $d(v) \leqq 16\pi \int_{-\pi}^{\pi} d\lambda \bar{f}^2(\lambda)$; further, for large values of v, approximately $d(v) = 2\pi \int_{-\pi}^{\pi} d\lambda \bar{f}^2(\lambda)$. One thus sees that in order to obtain bounds for $d(v)$ one must have a knowledge of the quantity

$$d = 2\pi \int_{-\pi}^{\pi} d\lambda \frac{f^2(\lambda)}{R^2(0)} = \frac{1}{R^2(0)} \sum_{v=-\infty}^{\infty} R^2(v). \tag{8.30}$$

In the study of both correlation analysis and spectral analysis of stationary time series it will be found that the quantity d arises frequently as information which one requires about the time series under consideration in order to carry out various statistical procedures. A satisfactory estimate of d from observations $\{X(t), t = 1, 2, \cdots, N\}$ is provided by

$$d_N = \frac{1}{2R_N^2(0)} \sum_{v=-(N-1)}^{N-1} R_N^2(v). \tag{8.31}$$

If one does not desire to compute the sample covariance function for all $v = 0, 1, \cdots, N$ then one may take, for any θ in $0 < \theta \leqq 1$,

$$d_{N,[\theta N]} = \frac{1}{(1 + 2\theta - \theta^2)R_N^2(0)} \sum_{v=-[\theta N]}^{[\theta N]} R_N^2(v) \tag{8.32}$$

as an estimate of d. The properties of the estimates d_N and $d_{N,[\theta N]}$ have been extensively investigated by Lomnicki and Zaremba [29]; among other things they show that d_N is a consistent estimate of d which in the case of a linear process has an asymptotic variance not dependent on the residuals $\{\eta(\alpha)\}$.

An alternate approach to the problem of investigating the mechanism generating a time series is to attempt to fit the time series by a finite parameter scheme (such as an autoregressive scheme or a moving average scheme). Here we consider only the problem of fitting an autoregressive scheme which has the most developed theory (for recent work on fitting moving average schemes, see Durbin [10]).

THEOREM 8E. *In order that a stationary time series $X(t)$ with covariance func-*

tion $R(v)$, satisfy the autoregressive scheme of order m,

$$X(t) = a_1X(t-1) + \cdots + a_mX(t-m) + \eta(t) \tag{8.33}$$

where $\eta(t)$ are a sequence of orthogonal random variables (with common variance σ^2) representing the innovation at time t so that

$$E[X(s)\eta(t)] = 0 \qquad \text{for } s < t \tag{8.34}$$

it is necessary and sufficient that the covariance function $R(v)$ satisfy the difference equation

$$R(u) = a_1R(u-1) + \cdots + a_mR(u-m) \qquad \text{for } u > 0 \tag{8.35}$$

while for $u = 0$

$$R(0) = a_1R(1) + \cdots + a_mR(m) + \sigma^2. \tag{8.36}$$

REMARK. Equations (8.35) are called the Yule-Walker equations, after G. Udny Yule and Sir Gilbert Walker who first obtained relations of this kind (see Wold [58], especially pp. 104–5 and pp. 140–146).

PROOF. Verify that (8.33) and (8.35) are each equivalent to the assertion that the minimum means square error linear predictor of $X(t)$, given its infinite past, depends only on the finite past $X(t-1), \cdots, X(t-m)$; in symbols, for all t

$$\begin{aligned} &E^*[X(t) \mid X(t-1), \cdots, X(t-m), \cdots] \\ &\qquad = a_1X(t-1) + \cdots + a_mX(t-m). \end{aligned} \tag{8.36}$$

We may use the fact that the covariance of a stationary autoregressive scheme satisfies the difference equation (8.35) to obtain expressions for the constants $a_1, \cdots, a_m$ in terms of correlations; (8.35) with $u = 1, \cdots, m$ yields m equations which may be written in matrix form

$$\begin{bmatrix} \rho(0) & \rho(1) & \cdots & \rho(m-1) \\ \rho(1) & \rho(0) & \cdots & \rho(m-2) \\ \vdots & & & \\ \rho(m-1) & \rho(m-2) & \cdots & \rho(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(m) \end{bmatrix}. \tag{8.37}$$

Consistent asymptotically normal estimates $a_1^{(N)}, \cdots, a_m^{(N)}$, of $a_1, \cdots, a_m$ respectively, may be obtained from observations $\{X(t), t = 1, 2, \cdots, N\}$ by forming consistent asymptotically normal estimates $\rho_N(v)$ of $\rho(v)$ and defining $a_1^{(N)}$ to be the solutions of

$$\begin{bmatrix} \rho_N(0) & \rho_N(1) & \cdots & \rho_N(m-1) \\ \rho_N(1) & \rho_N(0) & \cdots & \rho_N(m-2) \\ \vdots & \vdots & \cdots & \vdots \\ \rho_N(m-1) & \rho_N(m-2) & \cdots & \rho_N(0) \end{bmatrix} \begin{bmatrix} a_1^{(N)} \\ a_2^{(N)} \\ \vdots \\ a_m^{(N)} \end{bmatrix} = \begin{bmatrix} \rho_N(1) \\ \rho_N(2) \\ \vdots \\ \rho_N(m) \end{bmatrix}.$$

It may be shown, using standard techniques of large sample statistical theory,

that if the estmates $R_N(v)$ satisfy (8.21), then the estimates $a_i^{(N)}$ satisfy

$$E[\exp i\{u_1(N^{\frac{1}{2}})(a_1^{(N)} - a_1) + \cdots + u_m N^{\frac{1}{2}}(a_m^{(N)} - a_m)\}] \tag{8.39}$$
$$\rightarrow \exp\left[-\frac{1}{2}\sum_{i,j=1}^{m} u_i \sigma^2 C_{ij} u_j\right]$$

where $\{C_{ij}\}$ is the inverse matrix of the m by m matrix whose (i, j)th entry is $\rho(i - j)$. As an *estimate* $\{C_{ij}^{(N)}\}$ of $\{C_{ij}\}$ one may take the inverse matrix of $\{\rho_N(i - j)\}$, and as an estimate of σ^2 one may take

$$\sigma_N^2 = \frac{1}{N - m}\sum_{t=m+1}^{N}\{X(t) - a_1^{(N)}X(t-1) - \cdots - a_m^{(N)}X(t-m)\}^2. \tag{8.40}$$

In words, (8.39) says that the usual theorems of regression analysis apply asymptotically to the problem of estimating the autoregressive coefficients, even though the regression functions $X(t - 1), \cdots, X(t - m)$ represent lagged values of the observed time series $X(t)$. This fact was first shown by Mann and Wald [31] whose paper is a fundamental contribution to the theory of time series analysis.

To prove (8.39), we write (8.37) and (8.38) in alternate form as follows. Define $a_0 = a_0^{(N)} = 1$. Then, for $i = 1, 2, \cdots, m$

$$\sum_{j=0}^{m} a_j \rho(i - j) = 0 \tag{8.37'}$$

$$\sum_{j=0}^{m} a_j^{(N)} \rho_N(i - j) = 0. \tag{8.38'}$$

Therefore for $i = 1, \cdots, m$

$$\sum_{j=0}^{m} \rho_N(i - j)\{a_j^{(N)} - a_j\} = \sum_{j=0}^{m} a_j\{\rho(i - j) - \rho_N(i - j)\}. \tag{8.41}$$

From (8.41) one may deduce (8.39).

EXAMPLE. Let us write out the foregoing formulas for the case of an autoregressive scheme of order 2. Then (8.38) may be written

$$\begin{aligned} a_1^{(N)} \quad &+ a_2^{(N)}\rho_N(1) = \rho_N(1) \\ a_1^{(N)}\rho_N(1) \; &+ a_2^{(N)} \quad = \rho_N(2). \end{aligned} \tag{8.42}$$

The estimates $a_1^{(N)}$ and $a_2^{(N)}$ are explicitly given by

$$\begin{aligned} a_1^{(N)} &= \frac{\rho_N(1)\{1 - \rho_N(2)\}}{1 - \rho_N^2(1)} \\ a_2^{(N)} &= \frac{\rho_N(2) - \rho_N^2(1)}{1 - \rho_N^2(1)}. \end{aligned} \tag{8.43}$$

The estimated covariance matrix of $\{\text{Cov}\,[a_i^{(N)}, a_j^{(N)}]\}$ is given by

$$\{\text{Cov}\,[a_i^{(N)}, a_j^{(N)}]\}_{\text{est}} = \frac{1}{N}\sigma_N^2 \begin{bmatrix} 1 & \rho_N(1) \\ \rho_N(1) & 1 \end{bmatrix}^{-1} \tag{8.44}$$

To test the null hypothesis that the time series obeys an autoregressive scheme or order 1 against the alternative hypothesis that it obeys an autoregressive scheme of order 2 one uses the statistic

$$\delta = \frac{|a_2^{(N)}|^2}{\text{Var}\,[a_2^{(N)}]} = \frac{N\{\rho_N(2) - \rho_N^2(1)\}^2}{\sigma_N^2\{1 - \rho_N^2(1)\}} \tag{8.45}$$

which under the null hypothesis is distributed as χ^2 with 1 degree of freedom. One may similarly give a test of the null hypothesis that the time series obeys an autoregressive scheme of order q' against the alternative hypothesis that it obeys an autoregressive scheme of order q (greater than q').

For an excellent review of both the small and large sample theory of goodness of fit tests for autoregressive schemes, we refer the reader to the monograph by E. J. Hannan [17]. For references to recent work on explosive stochastic difference equations, see Rao [44a].

REFERENCES

[1] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, Vol. 68 (1950), pp. 337–404.

[2] M. S. Bartlett, "On the theoretical specification and sampling properties of autocorrelated time series," *J. Roy. Stat. Soc., Suppl.*, Vol. 8 (1946), pp. 27–41.

[3] M. S. Bartlett, *An Introduction to Stochastic Processes*, Cambridge, 1955.

[4] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra from the Point of View of Communications Engineering*, New York, Dover Publications, 1959.

[5] H. W. Bode and C. E. Shannon, "A simplified derivation of linear least squares smoothing and prediction theory," *Proc. I.R.E.*, Vol. 38 (1950), pp. 417–425.

[6] C. E. P. Brooks and N. Carruthers, *Handbook of Statistical Methods in Meteorology*, London, H. M. Stationary Office, 1953.

[7] K. A. Bush and I. Olkin, "Extrema of quadratic forms with applications to statistics," *Biometrika*, Vol. 46 (1959), pp. 483–486.

[8] Lola S. Deming, "Selected bibliography of statistical literature, 1930–1957: II. Time series," *J. Res. Nat. Bur. Standards*, Vol. 64B (1960), pp. 69–76.

[9] J. L. Doob, *Stochastic Processes*, New York, John Wiley and Sons, 1953.

[10] J. Durbin, "Efficient estimation of parameters in moving average models," *Biometrika*, Vol. 46 (1959), pp. 306–316.

[11] R. A. Fisher, "Tests of significance in harmonic analysis," *Proc. Roy. Soc.*, Ser. A. Vol. 125 (1929), p. 54.

[12] U. Grenander, "Stochastic processes and statistical inference," *Arkiv fur Mathematik*, Vol. 1 (1950), pp. 195–277.

[13] U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, New York, John Wiley and Sons, 1957.

[14] U. Grenander and G. Szego, *Toeplitz Forms and their Applications*, Berkeley, University of California Press, 1958.

[15] T. N. E. Greville, "Some applications of the pseudo inverse of a matrix," *SIAM Review*, Vol. 2 (1960), pp. 15–22.

[16] P. Halmos, *Introduction to Hilbert Space*, Chelsea, New York, 1951.

[17] E. J. Hannan, *Time Series Analysis*, Methuen, London, 1960.

[18] H. O. HARTLEY, "Tests of significance in harmonic analysis," *Biometrika*, Vol. 36 (1949), pp. 194–201.
[19] M. P. HEBLE, "Some problems in time series analysis," *J. Math. and Mechanics*, Vol. 9 (1960), pp. 951–987.
[20] G. M. JENKINS, "General considerations in estimation of spectra," *Technometrics*, Vol. 3 (1961), pp. 133–166.
[21] K. KARHUNEN, "Über linear methoden in der Wahrscheinlichkeitsrechnung," *Ann. Acad. Sci. Fennicae*, I. Math.-Physica, No. 37 (1947).
[22] M. G. KENDALL, "On the analysis of oscillatory time series," *J. Roy. Stat. Soc.*, Vol. 106 (1945), p. 91.
[23] M. G. KENDALL, *Advanced Theory of Statistics*, Vol. II, London, Charles Griffin, 1946.
[24] A. KHINTCHINE, "Korrelationstheorie der stationaren stochastischen Prozesse," *Math. Ann.*, Vol. 109 (1933), pp. 604–615.
[25] A. KOLMOGOROV, "Interpolation and extrapolation," *Bull. Acad. Sci. U.R.S.S.*, Ser. Math., Vol. 5 (1941), pp. 3–14.
[26] J. H. LANNING AND R. H. BATTIN, *Random Processes in Automatic Control*, New York, McGraw Hill Book Co., 1956.
[27] M. LOÈVE, "Fonctions aléatoires du second ordre," Suplement to P. Lévy, *Processus Stochastiques et Mouvement Brownien*, Paris, Gauthier-Villars, 1948.
[28] M. LOÈVE, *Probability Theory*, New York, D. Van Nostrand, 1955.
[29] Z. A. LOMNICKI AND S. K. ZAREMBA, "On some moments and distributions occurring in the theory of linear stochastic processes," *Monatshefte für Mathematik*, Part I, Vol. 61 (1957), pp. 318–358; Part II, Vol. 63 (1959), pp. 128–168.
[30] M. S. LONGUET-HIGGINS, "On the intervals between successive zeros of a random function," *Proc. Roy. Soc.*, Ser. A, Vol. 246 (1958), pp. 99–118.
[31] H. B. MANN AND A. WALD, "On the statistical treatment of linear stochastic difference equations," *Econometrica*, Vol. 11 (1943), pp. 173–220.
[32] D. MIDDLETON, *Statistical Communication Theory*, New York, McGraw Hill Book Co., 1960.
[33] P. A. P. MORAN, "Random processes in economic theory and analysis," *Sankhyā*, Vol. 21 (1959), pp. 99–126.
[34] J. NEYMAN, "Current problems of mathematical statistics," *Proceedings of the International Congress of Mathematicians*, Amsterdam, 1954, Volume 1.
[35] E. PARZEN, "A central limit theorem for multilinear stochastic processes," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 252–255.
[36] E. PARZEN, "On consistent estimates of the spectrum of a stationary time series," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 329–348.
[37] E. PARZEN, "Statistical inference on time series by Hilbert space methods, I," Department of Statistics, Stanford University, Technical Report No. 23, January 2, 1959.
[38] E. PARZEN, *Modern Probability Theory and its Applications*, John Wiley and Sons, 1960.
[39] E. PARZEN, "A new approach to the synthesis of optimal smoothing and prediction systems," Department of Statistics, Stanford University, Technical Report No. 34, July 15, 1960. (To be published in the *Proceedings of a Symposium on Optimization*, edited by R. Bellman.)
[40] E. PARZEN, "Regression analysis of continuous parameter time series, II," Department of Statistics, Stanford University, Technical Report No. 35, July 28, 1960. (To be published in the *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory*, edited by J. Neyman.)
[41] E. PARZEN, "Mathematical considerations in estimation of spectra," *Technometrics*, Vol. 3 (1961), pp. 167–190.
[42] E. PARZEN, "Estimating the covariance function of a stationary time series," to be published.

[42a] E. PARZEN AND G. R. HEXT, "A Bibliography of American Publications on Stochastic Processes and Time Series Analysis Published in the years 1900–1959 (Preliminary Report)," Stanford University Technical Report, 1961.

[43] V. S. PUGACHEV, "A method of solving the basic integral equation of statistical theory of optimum systems in finite form" (in Russian), *J. Applied Math. and Mechanics*, Vol. 23 (1959), pp. 3–14; English translation, pp. 1–16.

[44] V. S. PUGACHEV, *Theory of Random Functions and its Application to Automatic Control Problems* (in Russian), 2nd ed., State Publishing House of Theoretical Technical Literature (Gostekhizdat), Moscow, 1960.

[44a] M. M. RAO, "Consistency and limit distributions of estimators of parameters in explosive stochastic difference equations," *Ann. Math. Stat.*, Vol. 32 (1961), pp 195–218.

[45] M. ROSENBLATT, "Statistical analysis of stochastic processes with stationary residuals," *Probability and Statistics*, edited by U. Grenander, John Wiley and Sons, New York, 1959, pp. 246–273.

[46] H. SCHEFFÉ, *Analysis of Variance*, John Wiley and Sons, New York, 1959.

[47] A. SCHUSTER, "On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena," *Terr. Mag.*, Vol. 3 (1898), p. 13.

[48] A. SCHUSTER, "On the periodicities of sunspots," *Phil. Trans.*, Ser. A., Vol. 206 (1906), p. 69.

[49] M. M. SIDDIQUI, "On the inversion of the sample covariance matrix in a stationary autoregressive process," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 585–588.

[50] D. SLEPIAN, "First passage problem for a particular Gaussian process," *Ann. Math. Stat.*, Vol. 32 (1961), pp. 610–612.

[51] E. SLUTSKY, "The summation of random causes as the source of cyclic processes," *Econometrika*, Vol. 5 (1937), p. 105 (translation of a 1927 paper).

[52] C. T. STRIEBEL, "Efficient estimation of regression parameters for certain second order stationary processes," Ph.D. dissertation, University of California, 1960.

[53] J. W. TUKEY, "An introduction to the measurement of spectra," *Probability and Statistics*, edited by U. Grenander, John Wiley and Sons, New York, 1959, pp. 300–330.

[54] A. M. WALKER, "The asymptotic distribution of serial correlation coefficients for autoregressive processes with dependent residuals," *Proc. Camb. Phil. Soc.*, Vol. 50 (1954), pp. 60–64.

[55] ABBOTT S. WEINSTEIN, "Alternative definitions of the serial correlation coefficient in short autoregressive sequences," *J. Amer. Stat. Ass.*, Vol. 53 (1958), pp. 881–892.

[56] N. WIENER, "Generalized harmonic analysis," *Acta. Math.*, Vol. 55 (1930), p. 117.

[56a] N. WIENER, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley and Sons, New York, 1949.

[57] H. WOLD, *Demand Analysis*, John Wiley and Sons, New York, 1953.

[58] H. WOLD, *A Study in the Analysis of Stationary Time Series*, Uppsala, 2nd ed., 1954 (with an appendix by Peter Whittle).

[59] D. YLVISAKER, "On a class of covariance kernels admitting a power series expansion," to be published.

[60] G. U. YULE, "On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers," *Phil. Trans.*, Ser. A, Vol. 226 (1927), p. 267.

[61] L. A. ZADEH AND J. R. RAGAZZINI, "An extension of Wiener's theory of prediction," *J. Applied Physics*, Vol. 21 (1950), pp. 645–655.

[62] L. A. ZADEH AND J. R. RAGAZZINI, "Optimum filters for the detection of signals in noise," *Proc. I. R. E.*, Vol. 40 (1952), pp. 1223–1231.