

# A New Approach to the Synthesis of Optimal Smoothing and Prediction Systems<sup>†</sup>

EMANUEL PARZEN

## 1. Introduction

This chapter describes a new approach to a wide class of smoothing and prediction problems. The method can be applied to either stationary or nonstationary time series, with discrete or continuous parameters. It can easily be extended to time series observed in space-time and also to multiple time series, that is, those for which the observed value at each point of space-time is not a real number but a vector of real numbers.

Over the past few years I have been studying relationships between the theory of second-order stationary random functions, time series analysis, the theory of optimum design of communications and control systems, and classical regression analysis and analysis of variance. In the spring of 1957 I observed that reproducing-kernel Hilbert spaces provide a unified framework for these varied problems. The results obtained in 1957–1958 were theoretical elaborations of this idea, and were stated in a lengthy Stanford technical report [1] completed in the fall of 1958. Since then I have been concerned with developing examples and applications, well aware that the reproducing-kernel Hilbert space approach would be of no value unless it could provide new answers as well as old ones. It is hoped that the results presented here provide evidence that this approach is of value.

It may be of interest to relate this approach to one that is being

---

<sup>†</sup> Prepared with partial support of the Office of Naval Research.

Reprinted by permission from *Mathematical Optimization Techniques*, University of California Press, Berkeley, Calif., 1963, pp. 75–108

developed in the Soviet Union by V. S. Pugachev ([2]–[5]). Pugachev has in recent years advanced a point of view that he calls the *method of canonic representations of random functions*, for which in a recent article [5] he makes the following claim: “The results of this article, together with the results of [previous] papers, permit us to state that the method of canonic representations of random functions is the foundation of the modern statistical theory of optimum systems.” The methods to be presented in this chapter appear to provide a more powerful and elegant means of achieving in a unified manner the results that Pugachev has sought to unify by the method of canonic representations.

It may also be of interest to describe the standard approach to prediction and smoothing problems. The pioneering work of Wiener [6] and Kolmogorov [7] on prediction theory was concerned with a stationary time series observed over a semi-infinite interval of time, and sought predictors having 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 this integral equation. Simplified methods for solution of this equation in the practically important, special case of rational spectral density functions were given by Zadeh and Ragazzini [8] and Bode and Shannon [9]. Zadeh and Ragazzini [10] 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 treating prediction and smoothing problems involving a finite time of observation and nonstationary time series. The methods employed were either to reduce the solution of the problem to the solution of a suitable integral equation (generalization of the Wiener–Hopf equation) or to employ expansions (of Karhunen–Loève type) of the time series involved. In this chapter, we describe an approach to smoothing and prediction problems that may be called *coordinate free*, which, by the introduction of suitable coordinate systems, contains these previous approaches as special cases.

Finally, let us briefly outline the class of problems for which we shall give a unified, rigorous, and general treatment. A wide variety of problems concerning communication and control, or both (involving such diverse problems as the automatic tracking of moving 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, forecasting, the analysis of economic fluctuations, and the analysis of any kind of record representing observation over time), may be regarded as special cases of the following problem:

Let  $T$  denote a set of points on a time axis such that at each point  $t$  in  $T$  an observation has been made of a random variable  $X(t)$ . Given the observations  $\{X(t), t \in T\}$ , and a quantity  $Z$  related to the observation in a manner to be specified, one desires to form in an optimum manner estimates and tests of hypotheses about  $Z$  and various functions  $\psi(Z)$ .

This imprecisely formulated problem provides the general context in which to pose the following usual problems of communication and control.

*Prediction or extrapolation:* Observe the stochastic process  $X(t)$  over the interval  $s - T \leq t \leq s$ ; then predict  $X(s + \alpha)$  for any  $\alpha > 0$ . The length  $T$  of interval of observation may be finite or infinite. The optimum system yielding the predicted value of  $X(s + \alpha)$  is referred to as an optimum dynamic system if it provides estimates of  $X(s + \alpha)$  for all  $\alpha > 0$ .

*Smoothing or filtering:* Over the interval  $s - T \leq t \leq s$ , observe the sum  $X(t) = S(t) + N(t)$  of two stochastic processes or time series  $S(t)$  and  $N(t)$ , representing signal and noise respectively; then estimate  $S(t)$  for any value of  $t$  in  $s - T \leq t \leq s$ . The terminology "smoothing" derives from the fact that often the noise  $N(t)$  consists of very high-frequency components compared with the signal  $S(t)$ ; predicting  $S(t)$  can then be regarded as attempting to pass a smooth curve through a very wiggly record.

*Smoothing and prediction:* Observe  $S(t) + N(t)$  over  $s - T \leq t \leq s$ ; then predict  $S(s + \alpha)$  for any  $\alpha > 0$ .

*Parameter estimation:* Over an interval  $0 \leq t \leq T$ , observe  $S(t) + N(t)$ , where  $S(t)$  represents the trajectory (given by  $S(t) = x_0 + vt + at^2/2$ , say) of a moving object and  $N(t)$  represents errors of measurement; then estimate the velocity  $v$  and acceleration  $a$  of the object. More generally, estimate such quantities as  $S(t)$  and  $dS(t)/dt$  at any time  $t$  in  $0 \leq t \leq T$ , when the signal is of the form  $S(t) = \beta_1 w_1(t) + \dots + \beta_q w_q(t)$ .

*Signal extraction and detection:* Observe  $X(t) = A \cos \omega(t - \tau) + N(t)$  over an interval  $0 \leq t \leq T$ ; then estimate the parameters  $A$  and  $\tau$ , or test the hypothesis that  $A = 0$  against the hypothesis that  $A > 0$ . This problem is not explicitly treated in this chapter, although it could be handled by means of the tools described here.

## 2. A New Approach to Prediction Problems

Let us consider a stochastic process or time series  $\{X(t), t \in T\}$ , which is a family of random variables indexed by a parameter  $t$  varying in some index set  $T$ . Assume that each random variable has a finite second moment. Let

$$K(s, t) = E[X(s)X(t)] \quad (2.1)$$

be the *covariance kernel* of the time series. It might be thought more logical to call the function defined by (2.1) the *product moment kernel*, and reserve the name covariance kernel for the function defined by

$$K(s, t) = \text{Cov}[X(s), X(t)] = E[X(s)X(t)] - E[X(s)]E[X(t)]. \quad (2.2)$$

This terminology seems cumbersome, however, and is not adopted. We shall call the function defined by (2.2) the *proper covariance kernel*.

Let  $Z$  be a random variable with finite second moment for which one knows the cross-covariance function  $\rho_Z(\cdot)$ , defined by

$$\rho_Z(t) = E[ZX(t)], \quad t \text{ in } T. \quad (2.3)$$

A basic problem in statistical communication theory—which, as we shall see, is also basic to the study of the structure of time series—is that of *minimum mean-square error linear prediction*: Given a random variable  $Z$  with finite second moment, and a time series  $\{X(t), t \in T\}$ , find that random variable, linear in the observations, with smallest mean-square distance from  $Z$ . In other words, if we desire to predict the value of  $Z$  on the basis of having observed the values of the time series  $\{X(t), t \in T\}$ , one method might be to take that linear functional in the observations, denoted by  $E^*[Z|X(t), t \in T]$ , of which the mean-square error as a predictor is least.†

The existence and uniqueness of, and conditions characterizing, the best linear predictor are provided by the projection theorem of abstract Hilbert-space theory. (For proofs of the following assertions concerning Hilbert-space theory, see any suitable text, such as Halmos [13].)

By an abstract Hilbert space is meant a set  $H$  (with members  $u, v, \dots$ , that are usually called vectors or points) that possesses the following properties:

i.  $H$  is a linear space. Roughly speaking, this means that for any vector  $u$  and  $v$  in  $H$ , and real numbers  $a$ , there exist vectors, denoted by

† The symbol  $E^*$  is used to denote a predictor because, in the case of jointly normally distributed random variables, the least linear predictor  $E^*[Z|X(t), t \in T]$  coincides with the conditional expectation  $E[Z|X(t), t \in T]$ . For an elementary discussion of this fact, see Parzen [11], p. 387, or [12], Chap. 2.

$u + v$  and  $au$ , respectively, that 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 for every real number  $a$ ,

- a.  $(au, v) = a(u, v)$
- b.  $(u + v, w) = (u, w) + (v, w)$ ,
- c.  $(v, u) = (u, v)$ ,
- d.  $(u, u) > 0$  if  $u \neq 0$ .

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

The Hilbert space spanned by a time series  $\{X(t), t \in T\}$  is denoted by  $L_2(X(t), t \in T)$  and is defined as consisting of all random variables  $U$  that are either finite linear combinations of the random variables  $\{X(t), t \in T\}$ , or are limits of such finite linear combinations in the norm corresponding to the inner product defined on the space of square-integrable random variables by

$$(U, V) = E[UV]. \quad (2.4)$$

In words,  $L_2(X(t), t \in T)$  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  $v^*$  be the unique vector in  $M$  satisfying*

$$\|v^* - v\| = \min_{u \in M} \|u - v\| \quad (2.5)$$

is that

$$(v^*, u) = (v, u) \text{ for every } u \text{ in } M. \quad (2.6)$$

The vector  $v^*$  satisfying (2.5) is called the *projection of  $v$  onto  $M$* , and is also written  $E^*[v|M]$ .

In the case that  $M$  is the Hilbert space spanned by a family of vectors  $\{x(t), t \in T\}$  in  $H$ , we write  $E^*[v|x(t), t \in T]$  to denote the projection of  $v$  onto  $M$ . In this case, a necessary and sufficient condition that  $v^*$  satisfy (2.5) is that

$$(v^*, x(t)) = (v, x(t)) \text{ for every } t \in T. \quad (2.7)$$

We are now in a position to solve the problem of obtaining an explicit expression for the minimum mean-square error linear prediction  $E^*[Z|X(t), t \in T]$ . From (2.7) it follows that the optimum linear predictor is the unique random variable in  $L_2(X(t), t \in T)$  satisfying, for all  $t$  in  $T$ ,

$$E[E^*[Z|X(t), t \in T]X(t)] = E[ZX(t)]. \quad (2.8)$$

Equation (2.8) may look more familiar if we consider the special case of an interval  $T = \{t : a \leq t \leq b\}$ . If one writes, heuristically,

$$\int_a^b X(s)w(s) ds \quad (2.9)$$

to represent a random variable in  $L_2(X(t), t \in T)$ , then (2.8) states that the weighting function  $w^*(t)$  of the best linear predictor

$$E^*[Z|X(t), t \in T] = \int_a^b w^*(s)X(s) ds, \quad (2.10)$$

must satisfy the generalized Wiener-Hopf equation

$$\int_a^b w^*(s)K(s, t) ds = \rho_Z(t), \quad a \leq t \leq b. \quad (2.11)$$

There is an extensive literature [14], [15], [16] concerning the solution of the integral equation in (2.11). However, this literature is concerned with an unnecessarily difficult problem—one in which the very formulation of the problem makes it difficult to be rigorous. The integral equation in (2.11) has a solution only if one interprets  $w^*(s)$  as a generalized function including terms that are Dirac delta functions and derivatives of delta functions.

A simple reinterpretation of (2.11) avoids all of these difficulties. Let us not regard (2.11) as an integral equation for the weighting function  $w^*(s)$ ; rather, let us compare (2.10) and (2.11). These equations say that *if one can find a representation for the function  $\rho_Z(t)$  in terms of linear operations on the functions  $\{K(s, \cdot), s \in T\}$ , then the minimum mean-square error linear predictor  $E^*[Z|X(t), t \in T]$  can be written in terms of the corresponding linear operations on the time series*

$\{X(s), s \in T\}$ . 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 rather one of hunting for a linear representation of  $\rho_X(t)$  in terms of the covariance kernel  $K(s, t)$ . A general method of finding such representations will be discussed in the following sections. In this section we illustrate the ideas involved by considering several examples.

*Example 2A.* Consider a stationary time series  $X(t)$ , with covariance kernel

$$K(s, t) = Ce^{-\beta|t-s|}, \quad (2.12)$$

which we have observed over a finite interval of time,  $a \leq t \leq b$ . Suppose that we desire to predict  $X(b+c)$  for  $c > 0$ . Now, for  $a \leq t \leq b$ , we have

$$\rho(t) = E[X(t)X(b+c)] = Ce^{-\beta(b+c-t)} = e^{-\beta c}K(b, t). \quad (2.13)$$

In view of (2.13), by the interpretation of (2.10) and (2.11) just stated, it follows that

$$E^*[X(b+c) | X(t), a \leq t \leq b] = e^{-\beta c}X(b). \quad (2.14)$$

The present methods yield a simple proof of a widely quoted fact. Define a stationary time series  $X(t)$  with a continuous covariance function  $R(s-t) = E[X(s)X(t)]$  to be *Markov* if, for any real numbers  $a < b$  and  $c > 0$ , the least linear predictor of  $X(b+c)$ , given  $X(t)$  over the interval  $a \leq t \leq b$ , is a linear function of the most recent value  $X(b)$ ; in symbols,  $X(t)$  is Markov if

$$E^*[X(b+c) | X(t), a \leq t \leq b] = A(c)X(b) \quad (2.15)$$

for some constant  $A(c)$  depending only on  $c$ .

Let us now establish the following result:

**DOOB'S THEOREM:** Equation (2.15) holds if and only if, for some constants  $C$  and  $\beta$ ,

$$R(u) = Ce^{-\beta|u|}. \quad (2.16)$$

**PROOF:** From the fact that

$$\rho(t) = E[X(b+c)X(t)] = R(b-t+c),$$

it follows by the projection theorem that (2.15) holds if and only if, for every  $a < b$ ,  $c > 0$ , and  $t$  in  $a \leq t \leq b$ , we have

$$R(b-t+c) = A(c)R(b-t). \quad (2.17)$$

By (2.17) it follows that for every  $d \geq 0$  and  $c \geq 0$  we have

$$R(d + c) = A(c)R(d). \quad (2.18)$$

Letting  $d = 0$ , we obtain  $A(c) = R(c)$ ; consequently, for every  $c \geq 0$  and  $d \geq 0$ ,  $R(u)$  satisfies the equation

$$R(d + c) = R(d)R(c). \quad (2.19)$$

It is well known (see Parzen [11], p. 263) that a continuous even function  $R(u)$  satisfying (2.19) is of the form of (2.16).

*Example 2B. (Reinterpretation of the Karhunen-Loève expansion.)* Many writers on statistical communication theory (see [17], pp. 96, 244, 338-352, [18], and [19]) have made use of what is often called the Karhunen-Loève representation of a random function  $X(t)$  of second order. The results obtained are clarified when looked at from the present point of view.

The fundamental fact underlying the Karhunen-Loève expansion may be stated as follows:

**MERCER'S THEOREM.** *If  $\{\varphi_n(t), n = 1, 2, \dots\}$  denotes the sequence of normalized eigenfunctions and  $\{\lambda_n, n = 1, 2, \dots\}$  the sequence of corresponding nonnegative eigenvalues satisfying the relations*

$$\int_a^b K(s, t) \varphi_n(s) ds = \lambda_n \varphi_n(t), \quad a \leq t \leq b, \quad (2.20)$$

$$\int_a^b \varphi_m(t) \varphi_n(t) dt = \delta(m, n), \quad (2.21)$$

where  $\delta(m, n)$  is the Kronecker delta function, equal to 1 or 0 depending on whether  $m = n$  or  $m \neq n$ , then the kernel  $K(s, t)$  may be represented by the series

$$K(s, t) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(s) \varphi_n(t), \quad (2.22)$$

and this series converges absolutely and uniformly for  $a \leq s, t \leq b$ .

If we wish to predict the value of a random variable  $Z$  on the basis of the observed values  $X(t)$ ,  $a \leq t \leq b$ , we may write an explicit expression for the minimum mean-square error linear predictor as follows:

$$\begin{aligned} E^*[Z | X(t), a \leq t \leq b] &= \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \int_a^b \rho_Z(t) \varphi_n(t) dt \\ &\quad \times \int_a^b X(s) \varphi_n(s) ds. \end{aligned} \quad (2.23)$$

In order to prove the validity of (2.23), we need to prove that the infinite series is well defined and that it satisfies (2.8). Now

$$E \left[ \int_a^b X(s) \varphi_m(s) ds \int_a^b X(t) \varphi_n(t) dt \right] \\ = \int_a^b \int_a^b K(s, t) \varphi_m(s) \varphi_n(t) ds dt = \lambda_n \delta(m, n). \quad (2.24)$$

Therefore the mean square of the infinite series in (2.23) is equal to

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n} \left| \int_a^b \rho_Z(t) \varphi_n(t) dt \right|^2. \quad (2.25)$$

Consequently, a necessary and sufficient condition that the infinite series in (2.23) be well defined is that the infinite series in (2.25) be finite, which may be shown always to be the case. Next, we can show that (2.8) is satisfied by verifying that, for any  $t$  in  $a \leq t \leq b$ ,

$$E \left[ X(t) \left\{ \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \int_a^b \rho_Z(s) \varphi_n(s) ds \int_a^b X(u) \varphi_n(u) ds \right\} \right] \\ = \sum_{n=1}^{\infty} \varphi_n(t) \int_a^b \rho_Z(s) \varphi_n(s) ds = \rho_Z(t). \quad (2.26)$$

If it is permissible to interchange the processes of summation and integration in (2.23), then we may write

$$E^*[Z | X(t), a \leq t \leq b] = \int_a^b w^*(s) X(s) ds, \quad (2.27)$$

where

$$w^*(s) = \sum_{n=1}^{\infty} \varphi_n(s) \frac{1}{\lambda_n} \int_a^b \rho_Z(t) \varphi_n(t) dt. \quad (2.28)$$

The condition for the infinite series in (2.28) to be well defined is that

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n^2} \left| \int_a^b \rho_Z(t) \varphi_n(t) dt \right|^2 < \infty. \quad (2.29)$$

It can be shown that if (2.29) holds, then (2.27) is valid. Although (2.25) is always finite, however, (2.29) rarely holds. The optimal predictor is not usually of the form of (2.27). Thus we again see that it is not desirable to reduce prediction problems to the solution of the integral equation in (2.11).

*Example 2C. (The method of shaping filters.)* Another technique employed in statistical communication theory is the method of *shaping filters* (see Lanning and Battin [14]). Let  $X(t)$  be a stochastic process with covariance kernel  $K(s, t)$ . Let  $\eta(t)$  be a white-noise process, and let  $W(t, s)$  be a weighting function such that for every  $t$  we have

$$X(t) = \int_{-\infty}^t W(t, s)\eta(s) ds. \quad (2.30)$$

In words, the time series  $X(t)$  is represented as the response to a white-noise input of a system ("filter") described by a time-varying impulse-response function  $W(t, s)$ . If (2.30) holds, then  $W(t, s)$  is called a *shaping filter* for the time series  $X(t)$ . We now show how to use shaping filters to solve the prediction problem, given a time series  $X(t)$  that has been observed over a semi-infinite range,  $-\infty < t < b$ .

If (2.30) holds, and if the cross-covariance function  $\rho_Z(t)$  may be written, for a square-integrable function  $r(s)$ , as

$$\rho_Z(t) = \int_{-\infty}^t W(t, s)r(s) ds, \quad -\infty < t < b, \quad (2.31)$$

then

$$E^*[Z | X(t), -\infty < t \leq b] = \int_{-\infty}^b r(s)\eta(s) ds. \quad (2.32)$$

To prove (2.32), note that, for  $-\infty \leq t \leq b$ , we have

$$\begin{aligned} E\left[\int_{-\infty}^t W(t, s)\eta(s) ds \int_{-\infty}^b r(s)\eta(s) ds\right] \\ = \int_{-\infty}^t W(t, s)r(s) ds = \rho_Z(t). \end{aligned} \quad (2.33)$$

The expression given by (2.32) can be further simplified if we make the following reasonable assumptions about the shaping filter. Let  $L_t$  and  $M_t$  be differential operators of orders  $n$  and  $m$  respectively:

$$\begin{aligned} L_t &= \sum_{k=0}^n a_k(t) \frac{d^k}{dt^k}, \\ M_t &= \sum_{k=0}^m b_k(t) \frac{d^k}{dt^k}. \end{aligned} \quad (2.34)$$

Let  $H_L(t, s)$  and  $H_M(t, s)$  be the respective one-sided Green's functions characterized by the property that any sufficiently differentiable func-

tion  $f$  is given by

$$\begin{aligned} f(t) &= \int_{-\infty}^t H_L(t, s) L_s f(s) ds \\ &= \int_{-\infty}^t H_M(t, s) M_s f(s) ds. \end{aligned} \quad (2.35)$$

Suppose that the covariance kernel of  $X(t)$  may be written

$$K(t, s) = \int_{-\infty}^{\min(t, s)} M_t H_L(t, u) M_s H_L(s, u) du, \quad (2.36)$$

or, equivalently, that

$$X(t) = \int_{-\infty}^t M_t H_L(t, s) \eta(s) ds. \quad (2.37)$$

For an interesting discussion of how to find differential operators satisfying (2.36), see Batkov [20]. It may be shown that if (2.36) holds, then the right-hand side of (2.32) may be written in the form

$$\int_{-\infty}^b dt \int_{-\infty}^t L_t H_M(t, u) \rho_Z(u) du \int_{-\infty}^t L_t H_M(t, u) X(u) du. \quad (2.38)$$

In the particular case  $M_t \equiv 1$ , (2.38) reduces to

$$\int_{-\infty}^b dt \{L_t \rho_Z(t)\} \{L_t X(t)\}. \quad (2.39)$$

For the sake of rigor, it should be noted that in (2.38) and (2.39) the highest-order derivative of the observed time series  $X(t)$  may not exist, and we must then write  $dX^{(n-1)}(t)$  for  $X^{(n)}(t) dt$ .

### 3. General Solution of the Problems of Linear Prediction

It is possible to give a treatment of problems of prediction and smoothing that distinguishes between the statistical and analytical aspects of the problem. Such methods as that of expansions in eigenfunctions used in example 2B and that of shaping filters used in example 2C are merely analytical means of evaluating certain abstract quantities that can be defined without reference to these methods. The statistical problems of prediction and smoothing may be solved in terms of these abstract quantities once and for all. Indeed, the theory we shall now describe underlies the solution of many optimization problems; for

example, it includes as a special case the theory of generalized inverses of matrices (see Greville [21] for references to the history of the notion).

The basic tool in our theory is the notion of the *reproducing-kernel space* corresponding to a covariance kernel  $K$ .

**THEOREM 3.1.** (*Existence and uniqueness of the reproducing-kernel Hilbert space corresponding to a covariance function.*) Let  $\{X(t), t \in T\}$  be a time series with covariance kernel  $K(s, t)$  given by (2.1). Let  $H(K)$  consist of all functions  $h(\cdot)$  defined on  $T$  and of the form, for some  $U$  in  $L_2(X(t), t \in T)$ ,

$$h(t) = E[X(t)U], \quad \text{for all } t \in T. \quad (3.1)$$

On  $H(K)$  define an inner product by

$$(h, h)_K = E|U|^2. \quad (3.2)$$

Then  $H(K)$  is a Hilbert space. Further,  $H(K)$  possesses the following two properties: (a) for every  $t \in T$ ,

$$K(\cdot, t) \quad \text{belongs to } H(K), \quad (3.3)$$

where  $K(\cdot, t)$  is the function defined on  $T$  with value at  $s$  equal to  $K(s, t)$ ; (b) for every  $t$  in  $T$  and  $h(\cdot)$  in  $H(K)$ ,

$$h(t) = (h, K(\cdot, t))_K. \quad (3.4)$$

One calls (3.4) the *reproducing property* of the kernel  $K(s, t)$ . Since (3.4) holds, we call  $H(K)$  a reproducing-kernel Hilbert space, with reproducing kernel  $K$  (for the theory of such spaces, see [22]). The reproducing-kernel Hilbert space  $H(K)$  is uniquely determined by the conditions (3.3) and (3.4).

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

$$\int_{-\infty}^{\infty} f(s)\delta(s-t) ds = f(t). \quad (3.5)$$

Consequently, the kernel  $K(s, t) = \delta(s-t)$  satisfies (3.4). It does not satisfy (3.3), however, and therefore it is not truly a reproducing kernel.

**THEOREM 3.2.** (*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-to-one inner product pre-

serving linear mapping under which  $X(t)$  and  $K(\cdot, t)$  are mapped into one another. Denote by  $(h, X)_K$  the random variable in  $L_2(X(t), t \in T)$  that 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

$$\rho_Z(t) = E[ZX(t)],$$

then

$$E^*[Z | X(t), t \in T] = (\rho_Z, X)_K, \quad (3.6)$$

with mean-square error of prediction given by

$$E[|Z - E^*[Z | X(t), t \in T]|^2] = E|Z|^2 - (\rho_Z, \rho_Z)_K. \quad (3.7)$$

PROOF. The validity of Theorem 3.2 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, we have

$$E[(h, X)_K(g, X)_K] = (h, g)_K, \quad (3.8)$$

$$E[Z(h, X)_K] = (\rho_Z, h)_K, \quad (3.9)$$

where  $\rho_Z(t) = E[ZX(t)]$ . Now a random variable in  $L_2(X(t), t \in T)$  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 thus:

$$\begin{aligned} E[|(h, X)_K - Z|^2] &= E[(h, X)_K^2] + E[Z^2] - 2E[Z(h, X)_K] \\ &= E[Z^2] + (h, h)_K - 2(\rho_Z, h)_K \\ &= E[Z^2] - (\rho_Z, \rho_Z)_K + (h - \rho_Z, h - \rho_Z)_K. \end{aligned} \quad (3.10)$$

From (3.10) it is immediately seen 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 3.2 is thus complete.

Theorem 3.2 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 functions), 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.

*Example 3A. (Eigenfunction expansions.)* Let  $X(t)$ ,  $a \leq t \leq b$ , be a time series of which the covariance kernel  $K(s, t)$  has the eigenfunction

expansion (2.22). The corresponding reproducing-kernel Hilbert space consists of all square-integrable functions  $h(t)$  on the interval  $a \leq t \leq b$  such that

$$\int_a^b |h(t)|^2 dt = \sum_{n=1}^{\infty} \left| \int_a^b h(t) \varphi_n(t) dt \right|^2$$

and

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n} \left| \int_a^b h(t) \varphi_n(t) dt \right|^2 < \infty. \quad (3.11)$$

The reproducing-kernel inner product between two such functions is given by

$$(h, g)_K = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \int_a^b h(t) \varphi_n(t) dt \int_a^b g(t) \varphi_n(t) dt. \quad (3.12)$$

The random variable  $(h, X)_K$  in  $L_2(X(t), a \leq t \leq b)$  corresponding to  $h(\cdot)$  in  $H(K)$  under the mapping described in Theorem 3.2 is given by (3.12) with  $g$  replaced by  $X$ .

*Example 3B. (Autoregressive schemes.)* The reproducing-kernel Hilbert space and inner product corresponding to time series of the type described in example 2C can be determined; the reader may easily infer them from (2.32) and (2.38). Here let us consider a stationary time series  $X(t)$ , observed over a finite interval  $a \leq t \leq b$ , of the type that statisticians call an *autoregressive scheme*.

A continuous-parameter stationary time series  $X(t)$  is said to be an autoregressive scheme of order  $m$  if its covariance function may be written (see Doob [23], p. 542) as

$$R(s-t) = E[X(s)X(t)] = \int_{-\infty}^{\infty} \frac{e^{i(s-t)\omega}}{2\pi \left| \sum_{k=0}^m a_k(i\omega)^{m-k} \right|^2} d\omega, \quad (3.13)$$

where the polynomial

$$\sum_{k=0}^m a_k z^{m-k}$$

has no zeros in the right-hand half of the complex  $z$ -plane. It can be shown that, given observations of such a time series over a finite interval  $a \leq t \leq b$ , the corresponding reproducing-kernel Hilbert space contains all functions  $h(t)$  on  $a \leq t \leq b$  that are continuously dif-

ferentiable of order  $n$ . The reproducing-kernel inner product is given by

$$(h, g)_K = \int_a^b (L_t h)(L_t g) dt + \sum_{j,k=0}^m d_{j,k} h^{(j-1)}(a) g^{(k-1)}(a), \quad (3.14)$$

where

$$L_t h = \sum_{k=0}^m a_k h^{(m-k)}(t), \quad (3.15)$$

$$\{d_{j,k}\}^{-1} = \left\{ \frac{\partial^{j+k-2}}{\partial t^{j-1} \partial u^{k-1}} R(t-u) \Big|_{t=a, u=a} \right\}. \quad (3.16)$$

The first- and second-order autoregressive schemes are of particular importance.

A 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 with input a white noise  $\eta'(t)$  (the symbolic derivative of a process  $\eta(t)$  with independent stationary increments):

$$\frac{dX}{dt} + \beta X = \eta'(t). \quad (3.17)$$

It should be remarked that, from a mathematical point of view, (3.17) should be written as

$$dX(t) + \beta X(t) dt = d\eta(t). \quad (3.18)$$

Even then, in saying that  $X(t)$  satisfies (3.17) or (3.18) we mean that

$$X(t) = \int_{-\infty}^t H(t-s) d\eta(s), \quad (3.19)$$

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 time series  $X(t)$  is

$$R(t-u) = \frac{1}{2\beta} e^{-\beta|t-u|}. \quad (3.20)$$

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

$$(f, g) = \int_a^b (f' + \beta f)(g' + \beta g) dt + 2\beta f(a)g(a). \quad (3.21)$$

More generally, corresponding to the covariance function

$$K(s, t) = Ce^{-\beta|s-t|}, \quad (3.22)$$

the reproducing-kernel inner product is

$$\begin{aligned} (h, g)_K &= \frac{1}{2\beta C} \left\{ \int_a^b (h' + \beta h)(g' + \beta g) dt + 2\beta h(a)g(a) \right\} \\ &= \frac{1}{2\beta C} \int_a^b (h'g' + \beta^2 hg) dt + \frac{1}{2C} \{h(a)g(a) + h(b)g(b)\}. \end{aligned} \quad (3.23)$$

The random variable  $(h, X)_K$  in  $L_2(X(t), a \leq t \leq b)$ , corresponding to  $h(\cdot)$  in  $H(K)$ , may be written as

$$\begin{aligned} (h, X)_K &= \frac{1}{2\beta C} \left\{ \beta^2 \int_a^b h(t)X(t) dt + \int_a^b h'(t) dX(t) \right\} \\ &\quad + \frac{1}{2C} \{h(a)X(a) + h(b)X(b)\}. \end{aligned} \quad (3.24)$$

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 (3.24) makes sense. In the case that  $h(\cdot)$  is twice differentiable, one may integrate by parts and write

$$\int_a^b h'(t) dX(t) = h'(b)X(b) - h'(a)X(a) - \int_a^b X(t)h''(t) dt. \quad (3.25)$$

A 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 with input a white noise  $\eta'(t)$ :

$$\frac{d^2X}{dt^2} + 2\alpha \frac{dX}{dt} + \gamma^2 X = \eta'(t). \quad (3.26)$$

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

$$R(t-u) = \frac{e^{-\alpha|u-t|}}{4\alpha\gamma^2} \left\{ \cos \omega(u-t) + \frac{\alpha}{\omega} \sin \omega|u-t| \right\}. \quad (3.27)$$

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

$$\begin{aligned} (h, g)_K &= \int_a^b (h'' + 2\alpha h' + \gamma^2 h)(g'' + 2\alpha g' + \gamma^2 g) dt \\ &\quad + 4\alpha\gamma^2 h(a)g(a) + 4\alpha h'(a)g'(a). \end{aligned} \quad (3.28)$$

To write  $(h, X)_K$ , we use the same considerations as those in (3.24).

#### 4. General Solution of the Problem of Linear Smoothing (Regression Analysis)

Let  $\{X(t), t \in T\}$  be a time series of which the proper covariance kernel

$$K(s, t) = \text{Cov}[X(s), X(t)] \quad (4.1)$$

is known. The mean-value function,

$$m(t) = E[X(t)], \quad (4.2)$$

is only assumed to belong to a known class  $M$ . One case of particular importance is that in which  $M$  consists of all finite linear combinations of  $q$  known functions  $w_1(t), \dots, w_q(t)$ , so that the mean-value function is of the form

$$m(t) = \beta_1 w_1(t) + \dots + \beta_q w_q(t) \quad (4.3)$$

for unknowns  $\beta_1, \dots, \beta_q$  that are to be estimated.

In this section we consider the problem of estimating various functionals of the true mean-value function  $m(\cdot)$ ; in statistical theory, this is known as the problem of regression analysis of time series (see Parzen [24]). We seek estimates that (a) are linear in the observations  $\{X(t), t \in T\}$  in the sense that they belong to  $L_2(X(t), t \in T)$ , (b) are unbiased, in a sense to be defined, and (c) have minimum variance among all linear unbiased estimates.

**THEOREM 4.1.** (*General solution of the problem of minimum variance unbiased linear estimation.*) Let  $\{X(t), t \in T\}$  be a time series with known proper covariance kernel  $K(s, t)$ , and unknown mean-value function  $m(t)$  belonging to a known class  $M$  of functions. Let  $H(K)$  be the corresponding reproducing-kernel Hilbert space, and assume that  $M$  is a subset of  $H(K)$ .

i. Between  $L_2(X(t), t \in T)$  and  $H(K)$  there exists a one-to-one linear mapping with the following properties: for every  $t$  in  $T$ , and  $h$  and  $g$  in  $H(K)$ ,

$$(K(\cdot, t), X)_K = X(t), \quad (4.4)$$

$$E_m[(h, X)_K] = (h, m)_K \quad \text{for all } m \text{ in } M, \quad (4.5)$$

$$\text{Cov}[(h, X)_K, (g, X)_K] = (h, g)_K, \quad (4.6)$$

where  $(h, X)_K$  denotes the random variable in  $L_2(X(t), t \in T)$  that cor-

responds under the mapping to the function  $h(\cdot)$  in  $H(K)$ . 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.

ii. A random variable  $(h, X)_K$  in  $L_2(X(t), t \in T)$  is said to be an unbiased linear estimate of the value  $m(t)$  at a particular time  $t$  of the mean-value function  $m(\cdot)$  if

$$E_m[(h, X)_K] = (h, m)_K = m(t) \quad \text{for all } m \text{ in } M. \quad (4.7)$$

The uniformly minimum variance unbiased linear estimate  $m^*(t)$  of  $m(t)$  is given by

$$m^*(t) = (E^*[K(\cdot, t) | \bar{M}], X)_K, \quad (4.8)$$

in which  $\bar{M}$  is the smallest Hilbert subspace of  $H(K)$  containing  $M$ , and  $E^*[K(\cdot, t) | \bar{M}]$  is the projection onto  $\bar{M}$  of  $K(\cdot, t)$ .

iii. In the special case that  $\bar{M}$  is finite dimensional and is spanned by  $q$  functions  $w_1, \dots, w_q$  that are linearly independent as functions in  $H(K)$ , we can write explicitly

$$Wm^*(t) = - \begin{vmatrix} (w_1, w_1)_K \cdots (w_1, w_q)_K & (X, w_1)_K \\ \vdots & \vdots \\ (w_q, w_1)_K \cdots (w_q, w_q)_K & (X, w_q)_K \\ w_1(t) \cdots w_q(t) & 0 \end{vmatrix}, \quad (4.9)$$

$$W \text{Var} [m^*(t)] = - \begin{vmatrix} (w_1, w_1)_K \cdots (w_1, w_q)_K & w_1(t) \\ \vdots & \vdots \\ (w_q, w_1)_K \cdots (w_q, w_q)_K & w_q(t) \\ w_1(t) \cdots w_q(t) & 0 \end{vmatrix}, \quad (4.10)$$

where

$$W = \begin{vmatrix} (w_1, w_1)_K \cdots (w_1, w_q)_K \\ \vdots \\ (w_q, w_1)_K \cdots (w_q, w_q)_K \end{vmatrix}. \quad (4.11)$$

More generally, for any linear function  $\psi(\beta)$  of the parameters  $\beta_1, \dots, \beta_q$ ,

$$\psi(\beta) = \psi_1\beta_1 + \cdots + \psi_q\beta_q, \quad (4.12)$$

where the constants  $\psi_1, \dots, \psi_q$  are known, the minimum variance unbiased linear estimate of  $\psi(\cdot)$  is

$$\psi^* = \psi_1\beta_1^* + \cdots + \psi_q\beta_q^*, \quad (4.13)$$

where  $\beta_1^*, \dots, \beta_q^*$  are any solution of the set of normal equations

$$\begin{bmatrix} (w_1, w_1)_K & \dots & (w_1, w_q)_K \\ \vdots & & \vdots \\ (w_q, w_1)_K & \dots & (w_q, w_q)_K \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_q^* \end{bmatrix} = \begin{bmatrix} (w_1, X)_K \\ \vdots \\ (w_q, X)_K \end{bmatrix}. \quad (4.14)$$

In particular, if the true mean-value function  $m(\cdot)$  is of the form

$$m(t) = \beta w(t), \quad (4.15)$$

where  $w(\cdot)$  is known and  $\beta$  is a constant to be estimated, then

$$m^*(t) = \beta^* w(t), \quad \beta^* = \frac{(w, X)_K}{(w, w)_K}, \quad (4.16)$$

$$\text{Var} [m^*(t)] = \frac{1}{(w, w)_K}. \quad (4.17)$$

If the true mean-value function is of the form

$$m(t) = \beta_1 w_1(t) + \beta_2 w_2(t), \quad (4.18)$$

where  $w_i(\cdot)$  are known functions and  $\beta_1$  and  $\beta_2$  are constants to be estimated, then

$$m^*(t) = \beta_1^* w_1(t) + \beta_2^* w_2(t), \quad (4.19)$$

$$\text{Var} [m^*(t)] = W^{11} w_1^2(t) + 2W^{12} w_1(t)w_2(t) + W^{22} w_2^2(t). \quad (4.20)$$

In (4.19), we have

$$\begin{aligned} \beta_1^* &= W^{11}(w_1, X)_K + W^{12}(w_2, X)_K, \\ \beta_2^* &= W^{21}(w_1, X)_K + W^{22}(w_2, X)_K, \end{aligned} \quad (4.21)$$

where

$$\begin{aligned} W^{11} &= \frac{(w_2, w_2)_K}{W}, \\ W^{22} &= \frac{(w_1, w_1)_K}{W}, \end{aligned} \quad (4.22)$$

$$W^{12} = W^{21} = -\frac{(w_1, w_2)_K}{W},$$

$$W = (w_1, w_1)_K(w_2, w_2)_K - |(w_1, w_2)_K|^2.$$

To establish Theorem 4.1 there is no need to employ the method of Lagrange multipliers as so many writers do (see, for example, Lanning

and Battin [14], pp. 300–302); rather, we use the *projection* theorem. The minimum-variance unbiased linear estimate of  $m(t)$  may be characterized as the linear functional  $(h, X)_K$  that, among all linear functionals satisfying

$$E_m[(h, X)_K] = (h, m) = m(t) = (K(\cdot, t), m)_K \quad (4.23)$$

for all  $m$  in  $M$ , has minimum norm square

$$\|h\|_K^2 = \text{Var} [(h, X)_K]. \quad (4.24)$$

By the projection theorem, the function in  $H(K)$  having minimum norm among all functions satisfying the restraints (4.23) is  $E^*[K(\cdot, t) | \bar{M}]$ . Consequently, (4.8) has been proved. For a complete proof of Theorem 4.1, the reader is referred to [24].

*Example 4A.* To illustrate the use of the foregoing formulas, let us consider an example that has been treated by many authors. The statement of this problem is given by Lanning and Battin ([14], pp. 294, 303, 307): “Consider the problem of predicting a future position of a moving target by a system which receives target data, in the presence of noise, starting at  $t = 0$ .” Its position  $S(t)$  is an unknown linear function of time  $t$ ,

$$S(t) = \beta_1 + \beta_2 t, \quad (4.25)$$

where  $\beta_1$  and  $\beta_2$  are unknown constants; in Section 6 we consider the case in which  $\beta_1$  and  $\beta_2$  are random variables. The observed  $X(t)$  is assumed to be the sum of  $S(t)$  and a stationary random noise  $N(t)$ , with covariance function

$$R(u) = E[N(t)N(t+u)] = Ce^{-\beta|u|}. \quad (4.26)$$

It is desired to use observations  $X(t)$ ,  $0 \leq t \leq T$ , to estimate the particle's position  $S(t)$  at any given time  $t$ . Since  $S(t) = E[X(t)]$ , the problem of estimating  $S(t)$  is equivalent to the problem of estimating the mean-value function of an observed time series. Consequently, the minimum-variance unbiased linear estimate  $S^*(t)$  of the value of  $S(t)$  at a particular time  $t$  is given by the right-hand side of (4.19), with  $w_1(t) = 1$  and  $w_2(t) = t$ . The inner products appearing in (4.22) are explicitly given by means of (3.23) as follows:

$$(1, 1)_K = \frac{\beta T + 2}{2C},$$

$$(1, t)_K = \frac{\beta^2 T^2 + 2\beta T}{4C\beta},$$

$$(t, t)_K = \frac{\beta^3 T^3 + 3\beta^2 T^2 + 3\beta T}{6C\beta^2}, \quad (4.27)$$

$$W = (1, 1)_K(t, t)_K - (1, t)_K^2 = \frac{(\beta T)^4 + 8(\beta T)^3 + 24(\beta T)^2 + 24(\beta T)}{48C^2\beta^2}.$$

The variance of the estimate  $S^*(t)$  is given by the right-hand side of (4.20).

If the time series  $X(t)$  is assumed to be normal (or Gaussian), or if 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  $m(t)$  as follows. Given a confidence level  $\alpha$ , let  $K_q(\alpha)$  denote the  $\alpha$  percentile of the  $\chi^2$  distribution with  $q$  degrees of freedom; in symbols,

$$P[\chi_q^2 \geq K_q(\alpha)] = \alpha. \quad (4.28)$$

In particular, for  $q = 2$  and  $\alpha = 95$  per cent,  $K_q(\alpha)$  is approximately 6.

It can be shown that if the space  $M$  of possible mean-value functions has finite dimension  $q$ , then the interval

$$m^*(t) - \sqrt{K_q(\alpha)} \sigma[m^*(t)] \leq m(t) \leq m^*(t) + \sqrt{K_q(\alpha)} \sigma[m^*(t)], \quad (4.29)$$

for all  $t$  in  $-\infty < t < \infty$ , is a simultaneous confidence band for all values of the mean-value function with a level of significance not less than  $\alpha$ ; that is, if  $m(\cdot)$  is the true mean-value function, then (4.29) holds with a probability greater than or equal to  $\alpha$ .

## 5. Iterative Evaluation of Reproducing-Kernel Inner Products

In this section we give an iterative method of evaluating the reproducing-kernel inner product  $(h, h)_K$  and corresponding random variable  $(h, X)_K$  that makes possible the approximate synthesis of an optimum linear communication or control system in the presence of noise for which the covariance kernel  $K$  can be of any form and can be known either analytically or numerically. The method to be described is a gradient method related to the method of steepest descent. For a general discussion of the role of such methods in solving integral equations, see Kantorovich ([25], Chap. III), and in solving partial differential equations and algebraic linear equations, see Forsythe and Wasow ([26], Sec. 2).

Let  $K(s, t)$  be a covariance kernel, defined for  $a \leq s, t \leq b$ . Let  $H(K)$  be the corresponding reproducing-kernel Hilbert space. Let  $C(a, b)$  be the space of continuous functions on the interval  $a$  to  $b$ .

For a given function  $h$  in  $H(K)$ , it is of interest to develop methods of generating sequences  $\{H_n\}$  of functions in  $C(a, b)$  having the properties that

$$\lim_{n \rightarrow \infty} E[|(X, h)_K - \int_a^b H_n(t)X(t) dt|^2] = 0, \quad (5.1)$$

$$(h, h)_K = \lim_{n \rightarrow \infty} \int_a^b \int_a^b H_n(s)K(s, t)H_n(t) ds dt. \quad (5.2)$$

It is easily shown that sequences  $\{H_n\}$  satisfying (5.1) and (5.2) exist. As in example 2B, let values  $\lambda_n$  be the eigenvalues (arranged in decreasing order,  $\lambda_1 \geq \lambda_2 \geq \dots$ ) and let  $\varphi_n(\cdot)$  be the corresponding eigenfunctions of the kernel  $K(s, t)$ . Then a function  $h$  belongs to  $H(K)$  if and only if

$$\int_a^b |h(t)|^2 dt = \sum_{n=1}^{\infty} \left| \int_a^b h(t)\varphi_n(t) dt \right|^2$$

and

$$(h, h)_K = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \left| \int_a^b h(t)\varphi_n(t) dt \right|^2 < \infty. \quad (5.3)$$

Consequently, define

$$H_n(t) = \sum_{k=1}^n \varphi_k(s) \frac{1}{\lambda_k} \int_a^b h(s)\varphi_k(s) ds. \quad (5.4)$$

Clearly  $H_n(\cdot)$  belongs to  $C(a, b)$ .

It may be verified that

$$\int_a^b \int_a^b H_n(s)K(s, t)H_n(t) ds dt = \sum_{k=1}^n \frac{1}{\lambda_k} \left| \int_a^b h(t)\varphi_k(t) dt \right|^2 \quad (5.5)$$

and

$$\int_a^b H_n(t)X(t) dt = \sum_{k=1}^n \frac{1}{\lambda_k} \int_a^b h(s)\varphi_k(s) ds \int_a^b X(t)\varphi_k(t) dt. \quad (5.6)$$

Therefore the sequence defined by (5.4) satisfies (5.1) and (5.2). It is not computationally convenient, however, to use (5.4), inasmuch as it involves the calculation of eigenvalues and eigenfunctions.

Define a transformation  $T$  on functions  $H$  in  $C(a, b)$  as follows:

$$TH(t) = \int_a^b H(s)K(s, t) ds, \quad a \leq t \leq b. \quad (5.7)$$

It can be proved that

$$\int_a^b H(t)X(t) dt = (TH, X)_K, \quad (5.8)$$

$$\int_a^b \int_a^b H(s)K(s, t)H(t) ds dt = (TH, TH)_K. \quad (5.9)$$

Next, define a sequence of functions  $H_n$  as follows: Let  $\alpha$  be a constant to be specified. Let  $H_0(t) = 1$ , or some other function in  $C(a, b)$ . For  $n \geq 1$ , let†

$$H_{n+1} = H_n - \alpha(TH_n - h). \quad (5.10)$$

We claim that if  $\alpha$  is chosen in an interval specified by (5.18) or (5.21), then the sequence  $H_n$  defined by (5.10) satisfies (5.1) and (5.2). To prove this assertion it suffices to show that

$$E[|(h, X)_K - (TH_n, X)_K|^2] = \|(h - TH_n)_K\|^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (5.11)$$

From (5.10) we may write

$$\begin{aligned} TH_{n+1} - h &= (TH_n - h) - \alpha T(TH_n - h) \\ &= (I - \alpha T)(TH_n - h), \end{aligned} \quad (5.12)$$

where  $I$  is the identity operator,  $Ih(t) = h(t)$ . From (5.12) it follows that, for  $n \geq 0$ ,

$$TH_n - h = (I - \alpha T)^n(TH_0 - h). \quad (5.13)$$

We next note that for any function  $g$  in  $H(K)$ ,

$$g(t) = \sum_{n=1}^{\infty} \varphi_n(t) \int_a^b \varphi_n(s)g(s) dt, \quad (5.14)$$

$$Tg(t) = \sum_{n=1}^{\infty} \varphi_n(t)\lambda_n \int_a^b \varphi_n(s)g(s) ds, \quad (5.15)$$

$$\|(I - \alpha T)g\|_K^2 = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \left\{ \int_a^b \varphi_n(s)g(s) ds \right\}^2 \{1 - \alpha\lambda_n\}^2. \quad (5.16)$$

---

† Leonov gives an iterative procedure similar to the one given here in his very interesting paper [27], which he correctly describes as the first application of the methods of functional analysis to the problem of determining the weight function of an optimal system. Although he mentions the problem of establishing the convergence of the procedure, the proof he sketches does not seem to be satisfactory.

Defining  $g = TH_0 - h$  and  $\gamma_n = \int_a^b \varphi_n(s)g(s) ds$ , from (5.13) and (5.16) we have

$$\|TH_n - h\|_K^2 = \sum_{m=1}^{\infty} \frac{1}{\lambda_m} \gamma_m^2 \{1 - \alpha_m\}^{2n}. \quad (5.17)$$

Let  $\alpha$  be chosen so that, for every integer  $m$ ,

$$-1 < 1 - \alpha\lambda_m < 1 \quad \text{or} \quad 0 < \alpha < 2/\lambda_m. \quad (5.18)$$

If (5.18) holds, then for any integer  $M$

$$\|TH_n - h\|_K^2 \leq \sum_{m=1}^M \frac{1}{\lambda_m} \gamma_m^2 \{1 - \alpha\lambda_m\}^{2n} + \sum_{m>M} \frac{1}{\lambda_m} \gamma_m^2, \quad (5.19)$$

which tends to 0 as we first let  $n$  tend to  $\infty$ , and then let  $M$  tend to  $\infty$  [note that the last term in (5.19) is the remainder term of a convergent series]. We have thus shown that if (5.18) is satisfied, then (5.11) holds. Further, the procedure converges monotonically, in the sense that

$$\|TH_{n+1} - h\|_K \leq \|TH_n - h\|_K. \quad (5.20)$$

If  $M$  is a constant such that  $\max_m \lambda_m < M$ , then (5.18) is satisfied if we choose  $\alpha$  so that

$$0 < \alpha \leq 2/M. \quad (5.21)$$

A convenient choice for  $M$  is

$$M = \sum_{m=1}^{\infty} \lambda_m = \int_a^b K(t, t) dt. \quad (5.22)$$

It should be remarked that (5.19) implies that

$$\lim_{n \rightarrow \infty} \int_a^b |(TH_n - h)(t)|^2 dt = 0, \quad (5.23)$$

since, for any  $g$  in  $H(K)$ ,

$$\begin{aligned} |g(t)|^2 &\leq \|g\|_K^2 K(t, t), \\ \int_a^b |g(t)|^2 &\leq \|g\|_K^2 \int_a^b K(t, t) dt. \end{aligned} \quad (5.24)$$

The iterative method given by (5.10) undoubtedly does not converge very quickly. Other iterative methods (such as an analogue of the conjugate gradient method [28]) can be developed and should be studied.

## 6. Random Regression Coefficients

Let  $\{X(t), t \in T\}$  be a time series of the form

$$X(t) = m(t) + Y(t). \quad (6.1)$$

It is assumed that  $Y(t)$  is a time series with known mean-value and covariance functions:

$$E[Y(t)] = 0, \quad E[Y(s)Y(t)] = R_Y(s, t). \quad (6.2)$$

It is assumed that  $m(t)$  is of the form

$$m(t) = \beta_1 w_1(t) + \cdots + \beta_q w_q(t), \quad (6.3)$$

where the functions  $w_1, \cdots, w_q$  are known, and  $\beta_1, \cdots, \beta_q$  are *random variables* independent of  $\{Y(t), t \in T\}$  with known means

$$\mu_j = E[\beta_j], \quad j = 1, \cdots, q, \quad (6.4)$$

and covariance matrix  $\Gamma = \{\Gamma_{ij}\}$ , where, for  $i, j = 1, \cdots, q$ ,

$$\Gamma_{ij} = \text{Cov} [\beta_i, \beta_j]. \quad (6.5)$$

We call the foregoing set of assumptions *the case of random regression coefficients*.

The problem of estimating (or predicting) the value of  $m(t)$  under the assumption of random regression coefficients has been considered by Lanning and Battin ([14], pp. 305-309) and Bendat ([29], Chap. 9). We here consider the more general problem of estimating a parametric function

$$\psi(\beta) = \psi_1 \beta_1 + \cdots + \psi_q \beta_q. \quad (6.6)$$

Strictly speaking, the problem before us is one of pure prediction. The minimum mean-square error predictor of the random variable  $\psi(\beta)$ , given the observations  $X(t), t \in T$ , is the projection  $E^*[\psi(\beta) | X(t), t \in T]$ . Consequently, our aim in this section is to give an explicit formula for the projection.

One answer to this problem was given in Section 2, namely

$$E^*[\psi(\beta) | X(t), t \in T] = (\rho, X)_{R_X}, \quad (6.7)$$

where

$$R_X(s, t) = E[X(s)X(t)], \quad \rho(t) = E[\psi(\beta)X(t)]. \quad (6.8)$$

We easily verify that

$$\begin{aligned} R_X(s, t) &= E[m(s)m(t)] + E[Y(s)Y(t)] \\ &= \sum_{j,k=1}^q w_j(s)(\Gamma_{jk} + \mu_j\mu_k)w_k(t) + R_Y(s, t), \end{aligned} \quad (6.9)$$

$$\rho(t) = E[\psi(\beta)X(t)] = \sum_{j,k=1}^q \psi_j(\Gamma_{jk} + \mu_j\mu_k)w_k(t). \quad (6.10)$$

We now propose to obtain an expression for the best estimate of  $\psi(\beta)$  in terms of the reproducing-kernel inner product corresponding to  $R_Y$ , and the matrices

$$\Gamma = \{\Gamma_{ij}\}, \quad K = \{K_{ij}\}, \quad K_{ij} = (w_i, w_j)_{R_Y}. \quad (6.11)$$

**THEOREM 6.1.** *The minimum mean-square error linear predictor of*

$$\psi(\beta) = \psi_1\beta_1 + \cdots + \psi_q\beta_q, \quad (6.12)$$

*given the observations  $\{X(t), t \in T\}$ , is*

$$\psi(\beta^*) = \psi_1\beta_1^* + \cdots + \psi_q\beta_q^*, \quad (6.13)$$

where

$$\begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_q^* \end{bmatrix} = (\Gamma^{-1} + K)^{-1} \left\{ \begin{bmatrix} (w_1, X)_{R_Y} \\ \vdots \\ (w_q, X)_{R_Y} \end{bmatrix} + \Gamma^{-1} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \end{bmatrix} \right\}. \quad (6.14)$$

*The estimates  $\beta_1^*, \dots, \beta_q^*$  have covariance matrix*

$$\{\text{Cov} [\beta_j^*, \beta_k^*]\} = \Gamma K (\Gamma^{-1} + K)^{-1}, \quad (6.15)$$

*and mean-square error matrix*

$$\{E[(\beta_j^* - \beta_j)(\beta_k^* - \beta_k)]\} = (\Gamma^{-1} + K)^{-1}. \quad (6.16)$$

*Application.* To understand the meaning of Theorem 6.1, let us consider the case  $q = 1$ . We then observe that  $X(t) = \beta w(t) + Y(t)$ , where  $Y(t)$  satisfies (6.2),  $w(t)$  is a known function, and  $\beta$  is a random variable (independent of  $Y(t)$ ,  $t \in T$ ) with mean  $\mu$  and variance  $\sigma^2$ .

The minimum mean-square error linear predictor of  $\beta$  is

$$\beta^* = \frac{\frac{\mu}{\sigma^2} + (w, X)_R}{\frac{1}{\sigma^2} + (w, w)_R}, \quad (6.17)$$

$$\text{Var} [\beta^*] = \frac{\sigma^2(w, w)_R}{\frac{1}{\sigma^2} + (w, w)_R}, \quad (6.18)$$

$$E[|\beta^* - \beta|^2] = \text{Var} [\beta] - \text{Var} [\beta^*] = \left\{ \frac{1}{\sigma^2} + (w, w)_R \right\}^{-1}. \quad (6.19)$$

On the other hand, if  $\beta$  is assumed to be an unknown constant rather than a random variable, then the minimum mean-square error unbiased linear estimate of  $\beta$  is

$$\beta^* = \frac{(w, X)_R}{(w, w)_R}, \quad (6.20)$$

$$E_\beta |\beta^* - \beta|^2 = \text{Var}_\beta [\beta^*] = \frac{1}{(w, w)_R}. \quad (6.21)$$

One sees that for  $\mu = 0$  and  $\sigma$  very large, (6.17) and (6.20) yield approximately the same expression for  $\beta^*$ . This result was previously obtained by Lanning and Battin ([14], p. 309).

**PROOF OF THEOREM 6.1.** Let us write *tr* to denote *transpose*, and define vectors  $\mu$ ,  $\beta$ ,  $\beta^*$ ,  $w(t)$  in the obvious manner; for example,  $\psi^{\text{tr}} = (\psi_1, \dots, \psi_d)$ . To prove (6.13), it suffices to prove that for every  $t$  in  $T$  we have

$$E[\beta X(t)] = E[\beta^* X(t)]. \quad (6.22)$$

Let  $A$  be the second-moment matrix of  $\beta$ , defined by  $A = \Gamma + \mu\mu^{\text{tr}}$ . Clearly we have

$$E[\beta X(t)] = Aw(t).$$

To evaluate the right-hand side of (6.22), let us write

$$\beta^* = (\Gamma^{-1} + K)^{-1}V + \mu,$$

where  $V^{\text{tr}} = (V_1, \dots, V_q)$ ,  $V_j = (w_j, X - E[X]_{R^p})$ , and  $E[X]$  is the function of  $t$  defined by  $E[X](t) = \mu^{\text{tr}}w(t)$ . It may be verified that

$$E[VX(t)] = (K\Gamma + I)w(t) = (\Gamma^{-1} + K)\Gamma w(t),$$

$$E[\beta^*X(t)] = \Gamma w(t) + \mu\mu^{\text{tr}}w(t) = Aw(t).$$

The proof of (6.22) is complete. To prove (6.15), verify that

$$\begin{aligned} \{\text{Cov} [\beta_j^*, \beta_k^*]\} &= (\Gamma^{-1} + K)^{-1}E[VV^{\text{tr}}](\Gamma^{-1} + K)^{-1}, \\ E[VV^{\text{tr}}] &= (K\Gamma + I)K = (\Gamma^{-1} + K)\Gamma K. \end{aligned}$$

To prove (6.16), verify that

$$\begin{aligned} \{E[(\beta_j^* - \beta_j)(\beta_k^* - \beta_k)]\} &= \{\text{Cov} [\beta_j, \beta_k]\} - \{\text{Cov} [\beta_j^*, \beta_k^*]\} \\ &= \Gamma - \Gamma K(\Gamma^{-1} + K)^{-1} \\ &= \{\Gamma(\Gamma^{-1} + K) - \Gamma K\}(\Gamma^{-1} + K)^{-1} \\ &= (\Gamma^{-1} + K)^{-1}. \end{aligned}$$

## 7. Minimum-Variance Linear Unbiased Prediction

Let  $\{X(t), t \in T\}$  be a time series of which the proper covariance function,

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

is known. The mean-value function  $m(t) = E[X(t)]$  is known only to be a member of a class  $M$  of possible mean-value functions, where  $M$  is a subset of the reproducing-kernel Hilbert space  $H(K)$  corresponding to  $K$ . To make the discussion concrete we assume that  $M$  consists of all functions  $m(t)$  of the form

$$m(t) = \beta_1 w_1(t) + \dots + \beta_2 w_2(t), \quad (7.2)$$

where the functions  $w_1, \dots, w_2$  are known.

Let  $Z$  be a random variable for which we know the variance  $\text{Var} [Z]$  and the covariance

$$\rho_Z(t) = \text{Cov} [Z, X(t)]. \quad (7.3)$$

The mean of  $Z$  depends on the true mean-value function as follows:

$$E[Z] = (h, m)_K \quad \text{for every } m \text{ in } M, \quad (7.4)$$

for some  $h$  in  $H(K)$ . If  $M$  consists of all functions of the form (7.2), then

$$E_{\beta}[Z] = \psi_1\beta_1 + \cdots + \psi_q\beta_q \quad (7.5)$$

for some known constants  $\psi_1, \cdots, \psi_q$ .

One case of particular importance is  $Z = X(t_0)$ , where  $t_0$  does not belong to  $T$ ; then  $\psi_j = w_j(t_0)$  for  $j = 1, \cdots, q$ .

It is desired to predict  $Z$ , given the observations  $\{X(t), t \in T\}$ . Now if the means  $E[X(t)]$  and  $E[Z]$  were known, then the minimum variance linear predictor  $Z^*$  of  $Z$  would satisfy

$$Z^* - E[Z] = (\rho_Z, X - m)_K, \quad (7.6)$$

from which it follows that

$$Z^* = (\rho_Z, X)_K + \sum_{i=1}^q \beta_i \psi_i - \sum_{i=1}^q \beta_i (\rho_Z, w_i)_K. \quad (7.7)$$

One might think it plausible in the case of unknown means that the minimum-variance unbiased linear predictor is given by

$$Z^* = (\rho_Z, X)_K + \sum_{i=1}^q \beta_i^* \{\psi_i - (\rho_Z, w_i)_K\}, \quad (7.8)$$

where  $\beta_1^*, \cdots, \beta_q^*$  are any solution of the "normal equations" given in (4.14). We now show that this conjecture is correct.

**THEOREM 7.1.** *Let  $\{X(t), t \in T\}$  have known proper covariance kernel  $K$ , and unknown mean-value function  $m$  belonging to a subspace  $M$  of  $H(K)$ . Let  $Z$  be a random variable with cross-covariance function  $\rho_Z(t) = \text{Cov}[Z, X(t)]$ ; and let its mean, for each  $m$  in  $M$ , be given by  $E_m[Z] = (h, m)_K$ , where  $h$  belongs to  $H(K)$ . The minimum-variance linear unbiased predictor  $Z^*$  of  $Z$ , given the observations  $\{X(t), t \in T\}$ , is*

$$Z^* = (X, \rho_Z)_K + (X, E^*[h - \rho_Z | M])_K, \quad (7.9)$$

with mean-square error of prediction

$$E|Z^* - Z|^2 = \text{Var}[Z] - \|\rho_Z\|_K^2 + \|E^*[h - \rho_Z | M]\|_K^2. \quad (7.10)$$

**REMARK.** A linear estimate  $(X, g)_K$  is said to be an *unbiased linear predictor* of  $Z$ , if for all  $m$  in  $M$  we have

$$E_m[(X, g)_K] = (m, g)_K = (m, h)_K = E_m[Z]. \quad (7.11)$$

The notion of unbiased linear prediction was first considered by Dolph and Woodbury [30].

PROOF. The mean-square error of prediction of an unbiased linear estimate of  $Z$  is given, independently of  $m$ , by

$$\begin{aligned} E | Z - (X, g)_K |^2 &= \text{Var} [Z - (X, g)_K] \\ &= \text{Var} [Z] + \text{Var} [(X, g)_K] \\ &\quad - 2 \text{Cov} [Z, (X, g)_K]. \end{aligned} \quad (7.12)$$

It may be shown that  $\rho_Z$  belongs to  $H(K)$  and that

$$\text{Cov} [Z, (X, g)_K] = (\rho_Z, g)_K. \quad (7.13)$$

In view of (7.13), we can write

$$\begin{aligned} E | Z - (X, g)_K |^2 &= \text{Var} [Z] + (g, g)_K - 2(\rho_Z, g)_K \\ &= \text{Var} [Z] - \|\rho_Z\|_K^2 + \|g - \rho_Z\|_K^2. \end{aligned} \quad (7.14)$$

Letting  $g = \rho_Z + f$ , we see that the best predictor is given by  $Z^* = (X, \rho_Z + f)_K$ , where  $f$  is the function of minimum norm  $\|f\|_K$  satisfying the constraints

$$(m, f)_K = (m, h - \rho_Z)_K \quad \text{for all } m \text{ in } M. \quad (7.15)$$

It is clear that  $f = E^* [h - \rho_Z | M]$ . The proof of Theorem 7.1 is now complete.

Let us now exhibit an explicit formula for the best predictor  $X^*(t)$  of  $X(t)$ , for  $t$  not in  $T$ . From Theorem 7.1, it follows that if  $m(t)$  is of the form of (7.2), then

$$WX^*(t) = - \begin{vmatrix} (w_1, w_1)_K \cdots (w_1, w_q)_K & (X, w_1)_K \\ \vdots & \vdots \\ (w_q, w_1)_K \cdots (w_q, w_q)_K & (X, w_q)_K \\ d_1(t) & \cdots & d_q(t) & (X, K(\cdot, t))_K \end{vmatrix}, \quad (7.16)$$

$$WE | X^*(t) - X(t) |^2 = \begin{vmatrix} (w_1, w_1)_K \cdots (w_1, w_q)_K & d_1(t) \\ \vdots & \vdots \\ (w_q, w_1)_K \cdots (w_q, w_q)_K & d_q(t) \\ d_1(t) & \cdots & d_q(t) & d(t) \end{vmatrix}, \quad (7.17)$$

where

$$d_j(t) = w_j(t) - (w_j, K(\cdot, t))_K, \quad (7.18)$$

$$d(t) = K(t, t) - (K(\cdot, t), K(\cdot, t))_K, \quad (7.19)$$

$$W = \begin{vmatrix} (w_1, w_1)_K \cdots (w_1, w_q)_K \\ \vdots \\ (w_q, w_1)_K \cdots (w_q, w_q)_K \end{vmatrix}. \quad (7.20)$$

## 8. Decision Theoretic Extensions

The problems considered in the foregoing discussion 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 develop communication theory from the viewpoint of statistical decision theory or any other theory of statistical inference (see [31], [32], [33]). 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 communication problems, 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 can write a formula for the probability density functional of a stochastic process that is normal (Gaussian).

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

$$K(s, t) = \text{Cov} [X(s), X(t)] \quad (8.1)$$

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 measure induced by normal time series with the same covariance kernel  $K$  and with 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 nonnegative  $P_1$ -integrable function, denoted by  $dP_2/dP_1$ , such that, for every measurable set  $B$  of sample functions,

$$P_2(B) = \int_B \frac{dP_2}{dP_1} dP_1 + P_2(BN). \quad (8.2)$$

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 that 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 [24], Sec. 6), 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 (Parzen [24]). Let  $P_m$  be the probability measure induced on the space of sample functions of a time series  $\{X(t), t \in T\}$  with covariance kernel  $K$  and mean-value function  $m$ . Assume that either (a)  $T$  is countable or (b)  $T$  is a separable metric space,  $K$  is continuous, and the stochastic process  $\{X(t), t \in T\}$  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  $m \in H(K)$ , then the probability density functional of  $P_m$  with respect to  $P_0$  is given by

$$f(X, m) = \frac{dP_m}{dP_0} = \exp \left\{ (X, m)_K - \frac{1}{2} (m, m)_K \right\}. \quad (8.3)$$

Using the concrete formula for the probability density functional of a normal process provided by (8.3), we have no difficulty in applying the concepts of classical statistical methodology to problems of inference on normal time series.

## References

1. Parzen, E., *Statistical Inference on Time Series by Hilbert Space Methods*, I, Department of Statistics, Stanford University, Technical Report No. 23, January 2, 1959.
2. Pugachev, V. S., *Theory of Random Functions and Its Application to Automatic Control Problems* (Russian), State Publishing House of Theoretical Technical Literature (Gostekhizdat), Moscow, 1957.
3. Pugachev, V. S., "Application of Canonic Expansions of Random Functions in Determining an Optimum Linear System" (Russian), *Automation and Remote Control*, Vol. 17, 1956, pp. 489-499.
4. Pugachev, V. S., "Integral Canonic Representations of Random Functions and Their Application in Determining Optimum Linear Systems" (Russian), *Automation and Remote Control*, Vol. 18, 1957, pp. 971-984; English translation, pp. 1017-1031.
5. Pugachev, V. S., "A Method of Solving the Basic Integral Equation of Statistical Theory of Optimum Systems in Finite Form" (Russian), *J. Appl. Math. Mech.*, Vol. 23, 1959, pp. 3-14; English translation, pp. 1-16.
6. Wiener, N., *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley & Sons, Inc., New York, 1949.
7. Kolmogorov, A., "Interpolation and Extrapolation," *Bull. Acad. Sci. URSS. Sér. Math.*, Vol. 5, 1941, pp. 3-14.
8. Zadeh, L. A., and J. R. Ragazzini, "An Extension of Wiener's Theory of Prediction," *J. Appl. Phys.*, Vol. 21, 1950, pp. 645-655.
9. Bode, H. W., and C. E. Shannon, "A Simplified Derivation of Linear Least-Squares Smoothing and Prediction Theory," *Proc. IRE*, Vol. 38, 1950, pp. 417-425.

10. Zadeh, L. A., and J. R. Ragazzini, "Optimum Filters for the Detection of Signals in Noise," *Proc. IRE*, Vol. 40, 1952, pp. 1223-1231.
11. Parzen, E., *Modern Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1960.
12. Parzen, E., *Stochastic Processes*, Holden-Day, San Francisco, 1962.
13. Halmos, P., *Introduction to Hilbert Space*, Chelsea Publishing Co., New York, 1951.
14. Lanning, J. H., and R. H. Battin, *Random Processes in Automatic Control*, McGraw-Hill Book Company, Inc., New York, 1956.
15. Miller, K. S., and L. A. Zadeh, "Solution of an Integral Equation Occurring in the Theories of Prediction and Detection," *Trans. IRE*, PGIT-2, No. 2, June, 1956, pp. 72-76.
16. Shinbrot, M., "Optimization of Time-varying Linear Systems with Non-stationary Inputs," *Trans. ASME*, Vol. 80, 1958, pp. 457-462.
17. Davenport, W. B., and W. L. Root, *Introduction to the Theory of Random Signals and Noise*, McGraw-Hill Book Company, Inc., New York, 1958.
18. Davis, R. C., "On the Theory of Prediction of Non-stationary Stochastic Processes," *J. Appl. Phys.*, Vol. 23, 1952, pp. 1047-1053.
19. Grenander, U., "Stochastic Processes and Statistical Inference," *Ark. Mat.*, Vol. 1, 1950, pp. 195-277.
20. Batkov, A. M., "Generalization of the Shaping Filter Method To Include Non-stationary Random Processes" (Russian), *Automation and Remote Control* (Russian), Vol. 20, 1959, pp. 1081-1094; English translation, pp. 1049-1062.
21. Greville, T. N. E., "Some Applications of the Pseudoinverse of a Matrix," *SIAM Rev.*, Vol. 2, 1960, pp. 15-22.
22. Aronszajn, N., "Theory of Reproducing Kernels," *Trans. Amer. Math. Soc.*, Vol. 68, 1950, pp. 337-404.
23. Doob, J. L., *Stochastic Processes*, John Wiley & Sons, Inc., New York, 1953.
24. Parzen, E., "Regression Analysis of Continuous Parameter Time Series," in Jerzy Newman (ed.), *Proceedings of the Fourth Berkeley Symposium on Probability and Mathematical Statistics*, Vol. I (*Theory of Statistics*), University of California Press, Berkeley and Los Angeles, California, 1961, pp. 469-489.
25. Kantorovich, L. W., "Functional Analysis and Applied Mathematics" (Russian), *Uspehi Mat. Nauk*, Vol. 3, 1948, pp. 89-189; English translation issued by National Bureau of Standards, Los Angeles, 1952.
26. Forsythe, G. E., and W. Wasow, *Finite Difference Methods for Partial Differential Equations*, John Wiley & Sons, Inc., New York, 1960.
27. Leonov, P., "On an Approximate Method for Synthesizing Optimal Linear Systems for Separating Signal from Noise" (Russian), *Automation and Remote Control*, Vol. 20, 1959, pp. 1071-1080; English translation, pp. 1039-1048.
28. Hestenes, M. R., and E. Stiefel, "Method of Conjugate Gradients for Solving Linear Systems," *J. Res. Nat. Bur. Standards*, Vol. 49, 1952, pp. 409-436.
29. Bendat, J. S., *Principles and Applications of Random Noise Theory*, John Wiley & Sons, Inc., New York, 1958.

30. Dolph, C. L., and M. A. Woodbury, "On the Relation between Green's Functions and Covariances of Certain Stochastic Processes and Its Application to Unbiased Linear Prediction," *Trans. Amer. Math. Soc.*, Vol. 72, 1952, pp. 519-550.
31. Middleton, D., *Introduction to Statistical Communication Theory*, Part IV, McGraw-Hill Book Company, Inc., New York, 1960.
32. Pugachev, V. S., "Determination of an Optimum System Using a General Criterion" (Russian), *Automation and Remote Control*, Vol. 19, 1958, pp. 519-539; English translation, pp. 513-532.
33. Pugachev, V. S., "Optimum System Theory Using a General Bayes Criterion," *Trans. IRE*, PGIT-6 No. 1, March, 1960, pp. 4-7.