Statistical inference on time series by RKHS methods

Emanuel Parzen¹
State University of New York at Buffalo

Dedicated to the beloved memory of Alfred Rényi

The theory of time series is studied by probabilists (under such names as Gaussian processes and generalized processes), by statisticians (who are mainly concerned with modelling discrete parameter time series by finite parameter schemes), and by communication and control engineers (who are mainly concerned with the extraction and detection of signals in noise). The aim of this review is to outline the unifying role of reproducing kernel Hilbert spaces (RKHS) in the theory of time series. There are 13 sections (which are divided into an introduction and 4 chapters). The section headings are:

0. The role of RKHS in time series analysis.
1. Time series, means and covariances.
2. RKHS representations of time series.
3. Interpolation with minimum norm.
4. Filtering and parameter estimation.
5. A general model of linear statistical inference on time series.
6. RKHS solution of the zero mean linear statistical inference problem.
7. Equivalence between time series parameter estimation, control theory, and approximation theory.
8. Integral representation theorem.
12. Covariance kernels of normal probability density functionals.

¹ Prepared with the support of an Office of Naval Research contract at Stanford University and a Fellowship at the IBM Systems Research Institute, New York, N.Y. 10017.
0. **The role of RKHS in time series analysis.** Among the aspects of time series analysis for which RKHS seem to provide (simple, general, and elegant) formulations are:

1. Estimation of parameters of linear models.
2. Regression analysis and design of experiments.
3. Relations of time series analysis to approximation theory.
5. Minimum variance unbiased estimation.
7. Properties of the probability measures on linear spaces induced by Gaussian processes.
8. Limit theorems for stochastic processes.

It would take a book to explain the theory and applications of RKHS in time series analysis. My aim in these lectures can only be to provide an introduction, which in turn is inevitably biased by my own interest in questions of *stostatistics* (statistical inference on stochastic processes).

It seems to me fruitful to regard the theory of time series as composed of three broad areas, which we call **statistical theory**, **probability theory**: *structural* and **probability theory**: *distributional*.

In the **statistical theory** of time series, one imagines that an observed time series $Y(\cdot) = \{Y(t), t \in T\}$ has the representation

$$Y(t) = Z(t; \theta) + N(t)$$

where, for each $t$, $Z(t; \theta)$ is a random variable which is a function of a parameter $\theta$, which could itself be a stochastic process. One seeks to estimate, test hypotheses, or make a decision about the parameter $\theta$ from the data $Y(\cdot)$. Often the first step in statistical inference of a parameter $\theta$ is to define and obtain the probability density functional (called by statisticians the *likelihood* function) of a sample path of $Y(\cdot)$ given a fixed parameter value $\theta$. When $\theta$ is a random element, this is the same as the *conditional probability density functional* of $Y(\cdot)$ given $\theta$. These notions can be made precise, using the notion of a Radon–Nikodým derivative, and one is then able to treat the problem of statistical inference on time series within the framework of the theory of statistical inference originally developed for independent observations.

In the **probability theory**: *structural* of time series, one seeks to regard a time series $\{Y(t), t \in T\}$ as a function chosen, in accordance with a probability measure $P[\cdot]$, from a space $\Omega_T$ whose members are functions on an index set $T$. For each $t$, $Y(t)$ is a random variable (function) with domain $\Omega_T$ whose value [denoted $Y(t, \omega)$] at a point $\omega$ in $\Omega_T$ is defined by $Y(t, \omega) = \omega(t)$. One seeks to (i) determine the spaces $\Omega_T$ on whose "measurable" subsets the time series $Y(\cdot)$ determines a probability
measure, and (ii) characterize and represent the “measurable” functions $f(\omega)$ having prescribed properties, such as $f$ is linear in the sense that
\[ f(\omega_1 + \omega_2) = f(\omega_1) + f(\omega_2), \]
or $f$ is square integrable.

The **probability theory: distributional** of time series is concerned with evaluating (exactly or approximately) the distribution of random variables $f(\omega)$ which arise as functionals of a time series.

In all these areas of time series analysis (especially for normal processes) RKHS provide elegant and general formulations of results and proofs. One reason for the central role of RKHS is that a basic tool in all these areas is the theory of equivalence and singularity of normal measures (probability measures corresponding to normal processes), and this theory seems to me to be most simply expressed in terms of RKHS.

This review is divided into four chapters. Chapter 1 (§§ 1, 2, 3) defines a time series and introduces their RKHS representation. Chapter 2 (§§ 4, 5, 6, 7) is an introduction to the statistical theory of time series analysis, emphasizing recent work on the relations of this theory to optimization. Chapter 3 (§§ 8, 9, 10) surveys the main examples of RKHS. Chapter 4 (§§ 11, 12) very briefly states the main results of the theory of equivalence and singularity of normal measures.

As far as the mathematical theory of RKHS is concerned, it should be noted that credit for the notion of a reproducing kernel is due to Stefan Bergman who introduced the notion in his 1920 Berlin thesis [see Bergman (1950) for references to the extensive work of Bergman]. The Hilbert space theory of RKHS is due to Aronszajn [see Aronszajn (1950)], although related ideas appear in the work of Krein (1949). A more abstract theory of reproducing kernels has been developed by Laurent Schwartz (1964). Aside from the writing of the above authors, the main mathematical sources of information on RKHS are the research papers of Devinatz [see Devinatz (1959) for references to his work] and a text in German by Meschkowski (1962).

**Chapter 1. Time Series and RKHS**

1. **Time series, means and covariances.** In order to define what we mean by a time series, we must first define “stochastic process.” There are three main approaches one can take to defining the notion of a stochastic process.

   (1) One can define a stochastic process to be an *indexed collection of random variables* which we write \( \{ Y(t), t \in T \} \) or \( \{ Y(t, \cdot), t \in T \} \). For each \( t \)
(in the index set $T$) $Y(t)$ or $Y(t, \cdot)$ is a random variable (real valued measurable function) on a probability space $(\Omega, \mathcal{A}, P)$. The value of $Y(t)$ or $Y(t, \cdot)$ at a point $\omega$ in $\Omega$ is denoted $Y(t, \omega)$.

(2) One can define a stochastic process to be a probability measure on a function space whose members are functions (possibly generalized functions) on an index set $T$. More generally, one can define a stochastic process to be a probability space $(V, \mathcal{B}_V, P_V)$ where $V$ is a topological linear space (especially a Banach space or a Hilbert space), $\mathcal{B}_V$ is the sigma-field of topological Borel sets (the smallest sigma-field of subsets of $V$ containing all open sets), and $P_V$ is a probability measure with domain $\mathcal{B}_V$.

(3) One can define a stochastic process to be a Banach or Hilbert space valued random element $Y$; in symbols, $Y: (\Omega, \mathcal{A}, P) \rightarrow (V, \mathcal{B}_V)$ and $Y^{-1}(\mathcal{B}_V) \subset \mathcal{A}$, where $(\Omega, \mathcal{A}, P)$ is an abstract probability space, $V$ is a Banach or Hilbert space, and $\mathcal{B}_V$ is the sigma-field of topological Borel sets in $V$.

It is my view that, in developing the theory of time series, it is most convenient to begin with the first point of view towards stochastic processes, since it can be shown that a stochastic process defined from the second and third points of view can be transformed into an equivalent process defined as an indexed collection of random variables.

From the point of view of "data analysis" a stochastic process is usually regarded as a model of a set of observations of a measured quantity over a period of time; it is natural in this case to adopt the view that a stochastic process is a family of random variables. In problems of "signal extraction" or "parameter estimation" one seeks an estimator of a vector $\theta$ which represents an unknown signal or parameter; regarding $\theta$ as a random vector is often convenient (and leads to so-called Bayesian estimators of $\theta$). When $\theta$ is an infinite dimensional random vector it is a stochastic process.

We use the phrase "time series" to describe a stochastic process which is being discussed as a family of random variables which, further, are assumed to have finite second moments and are being studied from the point of view of the behavior of their mean $m$ and covariance $K$. The definition of mean and covariance of a stochastic process is given below [it depends on the point of view one has adopted in defining the process; compare (1) and (4.29)].

Formal definition. A time series is a family $Y(\cdot) = \{Y(t), t \in T\}$ of random variables with finite second moments. The mean value function $m$ and covariance kernel $K$ of a time series are functions on $T$ and $T \otimes T$ respectively defined by (for all $s, t$ in $T$)

\begin{align*}
(1) \quad m(t) &= E[Y(t)], \quad K(s, t) = \text{Cov}[Y(s), Y(t)].
\end{align*}

A time series is called normal if for every integer $n$, $n$ points $t_1, \ldots, t_n$
and real numbers \( u_1, \ldots, u_n \), the random variable \( \sum_{j=1}^{n} u_j Y(t_j) \) is normal, or equivalently
\[
E \left[ \exp \left( i \sum_{j=1}^{n} u_j Y(t_j) \right) \right] = \exp \left\{ i \sum_{j=1}^{n} u_j m(t_j) - \frac{1}{2} \sum_{j,k=1}^{n} u_j u_k K(t_j, t_k) \right\}.
\]

We will see that it is often convenient to think of a time series \( Y \) as indexed by a Hilbert space \( H^0 \), with inner product \((x, y)_{H^0}\), in the sense that the inner product \((Y, h)_{H^0}\) between \( Y \) and a fixed vector \( h \) in \( H^0 \) is not a true inner product but symbolizes a random variable indexed by \( h \). There are many Hilbert spaces for which such "random inner products" can be defined. The basic technique for defining such "random inner products" is to define a correspondence between elements in \( H^0 \) and random variables in the Hilbert space denoted
\[
H_Y = H(Y(t), t \in T) = L_2(Y(t), t \in T),
\]
called the Hilbert space spanned by the time series \( Y(\cdot) \), and defined as the smallest Hilbert space of random variables containing \( Y(t) \) for each \( t \in T \). A random variable \( U \) is called a simple linear functional of \( Y(\cdot) \) if it is of the form
\[
U = \sum_{j=1}^{n} c_j Y(t_j)
\]
for some integers \( n \), reals \( c_1, \ldots, c_n \), and indices \( t_1, \ldots, t_n \). It can be shown that every random variable in \( H_Y \) is either a simple linear functional of \( Y(\cdot) \) or is a limit in quadratic mean of simple linear functionals of \( Y(\cdot) \). Intuitively we think of \( H_Y \) as the space of all linear functionals of \( Y(\cdot) \).

2. RKHS representations of time series. The existence of a reproducing kernel Hilbert space (RKHS) representation of a time series was noted by Loève (1948): their systematic use for statistical inference was discussed by me in 1958 at the International Congress of Mathematicians and was first published in 1959 (reprinted in Parzen (1967), p. 253). Here we summarize the basic ideas and notation with emphasis on obtaining a representation of a random variable which is a linear functional of a stochastic process \( \{\theta(s), s \in S\} \).

A kernel \( Q \), with domain \( S \otimes S \), is called nonnegative definite if for every integer \( n \), points \( s_1, \ldots, s_n \) in \( S \), and real constants \( c_1, \ldots, c_n \)
\[
\sum_{i,j=1}^{n} c_i Q(s_i, s_j) c_j \geq 0.
\]
A kernel \( Q \) is called symmetric if \( Q(s_1, s_2) = Q(s_2, s_1) \).
If $Q$ is the covariance kernel of a time series, it is symmetric and non-negative definite.

According to a basic theorem of E. H. Moore and N. Aronszajn [see Aronszajn (1950)] every symmetric nonnegative definite kernel $Q$ possesses a unique reproducing kernel Hilbert space (RKHS) with reproducing kernel $Q$, denoted $H(Q)$ or $\text{RKHS}(Q)$, defined as follows: $H(Q)$ is a Hilbert space of functions $f$ on $S$ with the properties, for all $s \in S$ and $f \in H(Q)$,

(I) $Q(\cdot, s) \in H(Q)$,
(II) $f(s) = (f, Q(\cdot, s))_{H(Q)}$.

In (I), $Q(\cdot, s)$ is the function on $S$ with value at $s' \in S$ equal to $Q(s', s)$.

Let $\{\theta(s), s \in S\}$ be a normal time series with mean $\bar{\theta}$ and covariance $Q$. Then the process $\tilde{\theta}(s) = \theta(s) - \bar{\theta}(s)$ has zero mean and covariance kernel

$$Q(s_1, s_2) = E[\tilde{\theta}(s_1)\tilde{\theta}(s_2)].$$

It can be shown that (i) every function $f$ in $H(Q)$ can be represented

$$f(s) = E[\tilde{\theta}(s)U] \quad \text{for some unique } U \in L_2(\tilde{\theta}(s), s \in S)$$

where $U$ has zero mean and variance satisfying

$$\|f\|_{H(Q)}^2 = E[U^2];$$

(ii) under the assumption that $\theta \in H(Q)$, we can set up a one-one correspondence between $H(Q)$ and $L_2(\theta(s), s \in S)$ such that

$$\theta(s) \leftrightarrow Q(\cdot, s),$$

$$\sum c_i\theta(s_i) \leftrightarrow \sum c_i Q(\cdot, s_i),$$

$$U = \text{l.i.m.} \sum c_i^{(m)} \theta(s_i^{(m)}) \leftrightarrow f = \text{l.i.m.} \sum c_i^{(m)} Q(\cdot, s_i^{(m)})$$

where l.i.m. denotes "limit in mean square" and l.i.m. denotes "limit in norm."

The random variable $U \in L_2(\theta(s), s \in S)$ corresponding to $f \in H(Q)$ we denote by $(f, \theta)_{H(Q)}$ or $f^\sim$; we call the first notation a "congruence inner product" because (when $H(Q)$ is infinite dimensional) it is not an inner product between two functions in $H(Q)$ but rather is the value at $f$ of a congruence mapping on $H(Q)$ to $L_2(\theta(s), s \in S)$ which maps $Q(\cdot, s)$ into $\theta(s)$.

One can show that under the assumptions:

(i) $\theta$ is a time series with mean $\bar{\theta}$ and covariance $Q$,
(ii) $\bar{\theta} \in H(Q)$,

$$E[(f, \theta)_{H(Q)}] = (f, \bar{\theta})_{H(Q)};$$

$$\text{Cov}[(f_1, \theta)_{H(Q)}, (f_2, \theta)_{H(Q)}] = (f_1, f_2)_{H(Q)},$$

for every $f, f_1, f_2 \in H(Q)$. 
In terms of the congruence inner product any random variable \( Z(t) \) which is a linear function of \( \{ \theta(s), s \in S \} \) can be expressed

\[
Z(t) = (a(t, \cdot), \theta)_{H(Q)} \quad \text{for some } a(t, \cdot) \in H(Q).
\]

Note that \( a(t, \cdot) \) is a function with domain \( S \); when (5) holds we sometimes write

\[
Z(t) = (a(t, s), \theta(s))_{H(Q)}.
\]

When (5) holds, we call \( a(t, \cdot) \) the representor of \( Z(t) \); intuitively it is obtained from \( \{ Q(\cdot, s), s \in S \} \) by the same linear operations by which \( Z(t) \) is obtained from \( \{ \theta(s), s \in S \} \). As an example

\[
Z(t) = \int_S A(t, s) \theta(s) \, ds \quad \text{if and only if } a(t, \cdot) = \int_S A(t, s) Q(\cdot, s) \, ds.
\]

3. **Interpolation with minimum norm.** One aim of this review is to provide an introduction to recent work on the relations between time series, control theory, and approximation theory, stimulated by the work of Kimeldorf and Wahba (1968), (1969). An important link in this study is the use of RKHS to provide a convenient notation for describing the solution to the following problem of abstract Hilbert space theory which at first glance has no relation to time series.

**Problem.** Find the vector in a Hilbert space \( H^0 \) which has minimum norm among all vectors satisfying a prescribed set of linear constraints.

We call this problem the problem of interpolation with minimum norm.

Let \( H^0 \) be a real Hilbert space, \( T \) an index set, and \( \{ \Psi(t), t \in T \} \) a family of vectors in \( H^0 \). Let \( Y \) or \( Y(\cdot) \) denote a real valued function on \( T \), and

\[
C = \{ U \in H^0 : Y(t) = (\Psi(t), U)_{H^0} \text{ for all } t \text{ in } T \}.
\]

Find the element in \( C \), denoted \( \hat{U} \), satisfying

\[
\| \hat{U} \|^2_{H^0} = \min_{U \in C} \| U \|^2_{H^0}.
\]

To solve this problem, we denote by \( H_{\Psi} \), or \( H(\Psi(t), t \in T) \), the Hilbert subspace of \( H^0 \) spanned by the family \( \{ \Psi(t), t \in T \} \); it consists of all finite linear combinations (for some \( n \), reals \( c_1, \ldots, c_n \) and indices \( t_1, \ldots, t_n \))

\[
v = \sum_{i=1}^n c_i \Psi(t_i)
\]

and limits in norm of such finite linear combinations. One can verify that the vector in \( C \) with minimum norm is the unique vector \( \hat{U} \) in \( H_{\Psi} \) which is in \( C \). By the solution to the problem of interpolation with minimum norm we mean an explicit formula or algorithm for \( \hat{U} \).
When the index set $T$ is finite, say $T = \{1, 2, \ldots, n\}$, an expression for the unique vector $\hat{U}$ in the intersection of $C$ and $H_\Psi$ is obtained as follows: $\hat{U}$ is of the form

\begin{equation}
\hat{U} = \sum_{s=1}^{n} \lambda_s \Psi(s)
\end{equation}

and satisfies for all $t$

\begin{equation}
Y(t) = (\hat{U}, \Psi(t))_{H^0} = \sum_{s=1}^{n} \lambda_s K(s, t)
\end{equation}

defining

\begin{equation}
K(s, t) = (\Psi(s), \Psi(t))_{H^0}.
\end{equation}

We call $K$ the covariance kernel of the family of vectors $(\Psi(t), t \in T)$; it will play a central role in our discussion.

Define matrices

\[
\Psi = \begin{bmatrix}
\Psi(1) \\
\vdots \\
\Psi(n)
\end{bmatrix}, \quad
Y = \begin{bmatrix}
Y(1) \\
\vdots \\
Y(n)
\end{bmatrix}, \quad
K = \begin{bmatrix}
K(1, 1) & \cdots & K(1, n) \\
\vdots & \ddots & \vdots \\
K(n, 1) & \cdots & K(n, n)
\end{bmatrix}.
\]

Assume $K$ to have an inverse $K^{-1}$. From (1) and (2) we obtain

\begin{equation}
K\lambda = Y, \quad \lambda = K^{-1} Y, \quad \hat{U} = \Psi' \lambda = \Psi' K^{-1} Y.
\end{equation}

Further one can verify that

\begin{equation}
\|U\|^2_{H^0} = \lambda' K \lambda = Y' K^{-1} Y.
\end{equation}

To extend these results to an arbitrary index set, we introduce the RKHS $H(K)$ corresponding to the covariance kernel $K$. Define

\[
H(K) = \{ f \text{ on } T: f(s) = (\Psi(s), U)_{H^0}, \text{ all } s \in T, \text{ for some unique } U \text{ in } H_\Psi, \|f\|^2_{H(K)} = \|U\|^2_{H^0}\}.
\]

One can verify that $H(K)$ is a Hilbert space of functions on $T$, with the properties: first, for all $t \in T$

(I) \hspace{1cm} K(\cdot, t) \in H(K)

since $K(s, t) = (\Psi(s), \Psi(t))_{H^0}$, second, for all $t \in T$ and $f \in H(K)$

(II) \hspace{1cm} f(t) = (f, K(\cdot, t))_{H(K)}$

since by definition the last inner product equals $(U, \Psi(t))_H$ which equals $f(t)$.
Since
\[(6) \quad K(s, t) = (\Psi(s), \Psi(t))_{H^0} = (K(\cdot, s), K(\cdot, t))_{H(K)}\]
there is a congruence between \(H_\Psi\) and \(H(K)\) such that \(\Psi(t)\) and \(K(\cdot, t)\) are transformed into each other; in symbols
\[(7) \quad \Psi(t) \leftrightarrow K(\cdot, t).\]
The element in \(H_\Psi\) corresponding to a function \(f\) in \(H(K)\) under this congruence will be denoted
\[(8) \quad f^\sim \quad \text{or} \quad (f, \Psi)_{H(K)}^\sim.
We call \((f, \Psi)_{H(K)}^\sim\) a congruence inner product; it is not a true inner product, but represents the vector in \(H_\Psi\) which is the same "linear combination" of \(\{\Psi(t), t \in T\}\) that \(f\) is of \(\{K(\cdot, t), t \in T\}\).
In terms of the congruence inner product notation, the element \(\hat{U}\) in \(H_\Psi\) satisfying
\[(9) \quad (\hat{U}, \Psi(t))_{H^0} = Y(t), \quad t \in T,\]
is given by
\[(10) \quad \hat{U} = Y^\sim = (Y, \Psi)_{H(K)}^\sim
with norm squared
\[(11) \quad \|\hat{U}\|_{H^0}^2 = \|Y\|_{H(K)}^2 = (Y, Y)_{H(K)}.
We consider (10) to be a useful explicit solution to the minimum norm interpolation problem. The congruence inner product may seem to be merely a notation for the solution rather than an explicit solution; in our view it is an explicit solution since many techniques for evaluating congruence inner products are available.
Let us summarize the main properties of congruence inner products: for any \(f, g\) in \(H(K)\) and \(t\) in \(T\),
\[(12) \quad ((f, \Psi)_{H(K)}, \Psi(t))_{H^0} = (f, (\Psi(\cdot), \Psi(t))_{H^0})_{H(K)}
= (f, K(\cdot, t))_{H(K)} = f(t),\]
\[(13) \quad ((f, \Psi)_{H(K)}^\sim, (g, \Psi)_{H(K)}^\sim)_{H^0} = (f, g)_{H(K)}.
When the Hilbert space \(H^0\) is a function space, for great clarity we write \(\Psi(t, \cdot)\) for the elements in \(H^0\); then we are seeking \(\hat{U}(\cdot)\) of minimum \(H^0\)-norm in the set of \(U\) satisfying \((\Psi(t, \cdot), U(\cdot))_{H^0} = Y(t)\). The solution will then be written
\[(14) \quad \hat{U}(\cdot) = (Y(t), \Psi(t, \cdot))_{H((\Psi(s, \cdot), \Psi(t, \cdot))_{H^0})}.\]
4. Filtering and parameter estimation. The modern era of time series analysis has basically been concerned with developing methods for solving the filtering problem: given a time series \( \{ Y(t), t \in T \} \) which is of the form [called signal \( X(\cdot) \) plus noise \( N(\cdot) \)]
\[
Y(t) = X(t) + N(t), \quad 0 \leq t \leq T,
\]
compute, for each \( t \in T \),
\[
\hat{X}(t) = E^p[X(t)|Y(s), s \in T],
\]
the "wide sense" conditional mean of \( X(t) \) given the entire set of \( Y(\cdot) \) values.

The wide sense conditional mean is best defined as the solution to the following optimization problem: Let \( H_Y = H(Y(t), t \in T) \) be the Hilbert subspace (of \( L_2(\Omega, \mathcal{F}, P) \), the space of square integrable random variables) spanned by the family \( \{ Y(t), t \in T \} \). Find \( \hat{U} \) in \( H_Y \) such that
\[
E[|\hat{U} - X(t)|^2] = \min_{U \in H_Y} E[|U - X(t)|^2].
\]
We denote \( \hat{U} \) by \( \hat{X}(t) \) or by \( E^p[X(t)|Y(s), s \in T] \) where the superscript \( p \) connotes "projection in Hilbert space."

The vector \( \hat{X}(t) \) can be shown to be the solution of the normal equations
\[
E[\hat{X}(t)Y(s)] = E[X(t)Y(s)], \quad s \in T.
\]

From (4) we write
\[
\hat{X}(t) = (E[X(t)Y(s)], Y(s))_{H(E[Y(t_1)Y(t_2)])}.
\]
In words, define the covariance kernel on \( T \otimes T \)
\[
K_Y(t_1, t_2) = E[Y(t_1)Y(t_2)].
\]
Let \( H(K_Y) \) be the RKHS corresponding to \( K_Y \). Then for each \( t \in T \)
\[
K_{XY}(t, s) = E[X(t)Y(s)] \in H(K_Y)
\]
and
\[
\hat{X}(\cdot) = (K_{XY}(\cdot, s), Y(s))_{H(K_Y)}.
\]

Typically the signal and noise are assumed to be uncorrelated. Denote by \( R \) and \( K \) the covariance kernel of \( N(\cdot) \) and \( X(\cdot) \) respectively;
\[
R(t_1, t_2) = E[N(t_1)N(t_2)],
\]
\[
K(t_1, t_2) = E[X(t_1)X(t_2)].
\]
Then
\begin{equation}
K_Y = R + K, \quad K_{XY} = K
\end{equation}
and
\begin{equation}
\hat{X}(\cdot) = (K(\cdot, s), Y(s))_{H(R + K)}.
\end{equation}

It may make (12) look more familiar if we relate it to the Wiener Hopf equation. Suppose \( T = \{ t : a \leq t \leq b \} \) and we write very unrigorously
\begin{equation}
\hat{X}(t) = \int_a^b W(t, u)Y(u)\,du.
\end{equation}

The normal equations (4) become
\begin{equation}
\int_a^b W(t, u)\{K(u, s) + R(u, s)\}\,du = K(t, s), \quad a \leq s \leq b,
\end{equation}
which is to be solved for the weighting function \( W(t, u) \). To write (14) in operator form define an operator \( R + K \) on \( L_2(a, b) \) as follows: \((R + K)f\) is the function whose value at \( s \) is
\begin{equation}
\{(R + K)f\}(s) = \int_a^b f(u)\{K(u, s) + R(u, s)\}\,du.
\end{equation}

In this notation (14) can be written
\begin{equation}
\{(R + K)W(t, \cdot)\}(s) = K(t, s)
\end{equation}
with formal solution
\begin{equation}
W(t, u) = \{(R + K)^{-1}K(t, \cdot)\}(u).
\end{equation}

From (13) and (17) we have (still arguing nonrigorously)
\begin{equation}
\hat{X}(t) = \int_a^b Y(u)\{(R + K)^{-1}K(t, \cdot)\}(u)\,du.
\end{equation}

Equation (12) is a rigorous restatement of the heuristic equation (18).

There are two important extensions of the basic model (1) of signal plus noise:

1. **Parameterization of the signal.** In addition to extracting the signal \( X(\cdot) \) we desire to estimate some underlying "parameter process" \( U(\cdot) \) of which \( X(\cdot) \) is a linear functional in a sense to be made precise.
(2) Transformation of the signal. The observed series \( Y(t) \) is the sum of a process \( Z(t) \) and noise \( N(t) \),

\[
Y(t) = Z(t) + N(t),
\]

where \( Z(t) \) is a linear transformation of the signal process \( X(t) \). We call \( Z(t) \) a transformed signal process.

Three important examples of parameterization of a signal are:

(i) the regression model;

\[
X(t) = \sum_{j=1}^{q} \beta_j \phi_j(t),
\]

in which \( X(t) \) is an unknown linear combination of \( q \) known functions \( \phi_1(t), \ldots, \phi_q(t) \),

(ii) the inner product model

\[
X(t) = \langle \Psi(t), U \rangle_{H^0}, \quad t \in T,
\]

in which \( U \) is a random element taking values in a Hilbert space \( H^0 \), and \( \{\Psi(t), t \in T\} \) is a known family of vectors in \( H^0 \),

(iii) the stochastic inner product model

\[
X(t) = \langle \Psi(t), U \rangle_{H^0}, \quad t \in T,
\]

in which \( \{\Psi(t), t \in T\} \) is a known family of elements in a Hilbert space \( H^0 \), and where we use the notation \( (h, U)_{H^0} \) to denote the value at \( h \) (an element in \( H^0 \)) of a Hilbert space indexed stochastic process \( \{U(h), h \in H^0\} \) with finite second moments which is linear in the sense that (for all \( h, h_1, h_2 \) in \( H^0 \))

\[
U(h_1 + h_2) = U(h_1) + U(h_2) \quad \text{with probability one},
\]

and is continuous in the sense that the mean value function

\[
m_U(h) = E[U(h)] = E[(h, U)_{H^0}]
\]

and the covariance kernel

\[
Q_U(h_1, h_2) = \text{Cov}[U(h_1), U(h_2)] = \text{Cov}[(h_1, U)_{H^0}, (h_2, U)_{H^0}]
\]

are continuous functions of their arguments. The implications of these assumptions will be explored at the end of this section.

Three important examples of signal transformation are:

(i) instantaneous transformation, when \( X(t) \) is a random vector of dimension \( q \), \( Z(t) \) is a random vector of dimension \( r \), and

\[
Z(t) = H(t)X(t)
\]

where \( H(t) \) is a known \( r \) by \( q \) matrix;
(ii) integral transformation, when the signal \( \{X(\tau), \tau \in T_x\} \) is a process with index set \( T_x \) for which one can define the notion of stochastic integral, and

\[
Z(t) = \int_{T_x} A(t, \tau)X(\tau)\,d\tau
\]

where \( A \) is a known kernel with domain \( T \otimes T_x \);

(iii) RKHS transformation, where \( Z(t) \) is assumed only to be a random variable in \( H_x = H[X(\tau), \tau \in T_x] \), the Hilbert space of random variables spanned by the signal process \( \{X(\tau), \tau \in T_x\} \); then corresponding to \( Z(t) \) there is a function \( a(t, \cdot) \) in \( H(K_x) \), the RKHS corresponding to the covariance kernel \( K_x \) of \( X(\cdot) \), in terms of which we can write

\[
Z(t) = (a(t, \cdot), X)_{H(K_x)}.
\]

The RKHS transformation is the most general; other examples of signal transformation can always be represented as an RKHS transformation.

It should be noted that parameterization is in a rough sense the inverse of transformation; if \( Z(\cdot) \) is considered the signal process, then \( X(\cdot) \) is a parameterization of \( Z(\cdot) \). We still would employ a parameterization of \( X(\cdot) \), since for computational tractability it is usually necessary to parametrize a process \( X(\cdot) \) by a white noise process \( U(\cdot) \). We conclude this section by discussing the notion of white noise appropriate for our purposes.

Let \( H^0 \) be a Hilbert space, and let \( \{U(h), h \in H^0\} \) be an \( H^0 \)-indexed time series which is linear (in the sense that (23) holds) and continuous (in the sense that \( m_U(h) \) and \( Q_U(h_1, h_2) \) are continuous). Then there exists an element \( m \) in \( H^0 \) and a bounded linear operator \( Q : H^0 \to H^0 \) satisfying

\[
m_U(h) = (m, h)_{H^0}, \quad Q_U(h_1, h_2) = (Qh_1, h_2)_{H^0}.
\]

We call \( m \) the mean vector and \( Q \) the covariance operator of the time series \( U(\cdot) \).

We call \( U(\cdot) \) white noise with intensity \( \lambda \) if it has zero mean, \( m = 0 \), and covariance operator a multiple of the identity operator \( I \):

\[
Q = \lambda I.
\]

When \( \lambda = 1 \), we call \( U(\cdot) \) white noise.

Corresponding to a zero mean time series \( \{Y(t), t \in T\} \) with covariance kernel \( Q \) with RKHS \( H(Q) \), there is an \( H(Q) \)-indexed white noise

\[
\{(h, Y)_{H(Q)}, h \in H(Q)\}.
\]
5. A general model of linear statistical inference on time series. We seek to extract a signal process $X(\cdot)$ from an observed process $Y(\cdot)$ which is the sum of $Z(\cdot)$ and $N(\cdot)$, where $Z(\cdot)$ is a transformed signal process.

We assume a signal process $\{X(\tau), \tau \in T_X\}$ with the representation

\begin{equation}
X(\tau) = \sum_{j=1}^{q} \beta_j \phi_j(\tau) + (\Psi(\tau, \cdot), U)_{-H^0}, \quad \tau \in T_X,
\end{equation}

where $H^0$ is a Hilbert space of functions, $\phi_1, \ldots, \phi_q$ are known functions (satisfying conditions to be specified), $\{\Psi(\tau, \cdot), \tau \in T_X\}$ is a family of functions in $H^0$, and $U$ is $H^0$-valued white noise. We call $U$ the parameter process.

For the transformed signal process $\{Z(t), t \in T\}$ we assume the representation

\begin{equation}
Z(t) = \sum_{j=1}^{q} \tau_j A\phi_j(t) + (A\Psi(t, \cdot), U)_{-H^0}
\end{equation}

where $A\phi_1, \ldots, A\phi_q$ are known functions (intuitively $A\phi_j$ is the result of operating on $\phi_j$ with an operator $A$, but since we have not specified a space to which $\phi_j$ belong we cannot specify the sense in which $A$ is an operator in its own right), and $\{A\Psi(t, \cdot), t \in T\}$ is a family of functions in $H^0$.

The two kinds of parameters in the representation we have assumed for $X(\cdot)$ represent respectively the contribution of initial conditions and ongoing input. The parameters $\beta_1, \ldots, \beta_q$ represent the initial conditions, and the expression in $\beta_1, \ldots, \beta_q$ represent the contribution of initial conditions were there no additional input. The expression in $U$ represents the contribution of the continuous input were the initial conditions zero.

For an example of a process with the representation (1) see equation (8.17).

For the observed process $Y(\cdot)$ we assume

\[ Y(t) = Z(t) + N(t) \]

where the noise process $N(\cdot)$ is independent of the parameter process $U(\cdot)$, has zero mean, and covariance kernel $R$:

\begin{equation}
E[N(t)] = 0, \quad E[N(s)N(t)] = R(s, t).
\end{equation}

The mean value function and covariance kernel of $Y(\cdot)$, $X(\cdot)$, and $Z(\cdot)$ depend on the attitude we adopt towards the vector parameter

\[ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} \]

which can be non-Bayes or Bayes.
Non-Bayes attitude: $\beta$ is a vector of unknown constants to be estimated. Bayes attitude: $\beta$ is a random vector which is independent of $U$ and $N(\cdot, \cdot)$, has mean $\tilde{\beta}$ and covariance $Q$:

$$E[\beta] = \tilde{\beta}, \quad E[(\beta - \tilde{\beta})(\beta - \tilde{\beta})'] = Q.$$  

In the non-Bayes case, the mean value function and covariance kernel of $X(\cdot)$, $Z(\cdot)$, and $Y(\cdot)$ are as follows (for fixed values of the parameter $\beta$):

$$E_\rho[X(\tau)] = \sum_{j=1}^{q} \beta_j \phi_j(\tau),$$

$$\text{Cov}[X(\tau_1), X(\tau_2)] = (\Psi(\tau_1, \cdot), \Psi(\tau_2, \cdot))_{H^0},$$

$$E_\rho[Z(t)] = \sum_{j=1}^{q} \beta_j A\phi_j(t),$$

$$\text{Cov}[Z(t_1), Z(t_2)] = (A\Psi(t_1, \cdot), A\Psi(t_2, \cdot))_{H^0},$$

$$E_\rho[Y(t)] = \sum_{j=1}^{q} \beta_j A\phi_j(t),$$

$$\text{Cov}[Y(t_1), Y(t_2)] = (A\Psi(t_1, \cdot), A\Psi(t_2, \cdot))_{H^0} + R(t_1, t_2).$$

The subscript $\beta$ on the expectation operator indicates that the probability measure depends on the value of the parameter $\beta$.

In the Bayes case the mean value function and covariance kernel of $X(\cdot)$, $Z(\cdot)$, and $Y(\cdot)$ are as follows:

$$E[X(\tau)] = \sum_{j=1}^{q} \bar{\beta}_j \phi_j(\tau),$$

$$\text{Cov}[X(\tau_1), X(\tau_2)] = \sum_{i,j=1}^{q} \phi_j(\tau_1)Q_{ij}\phi_i(\tau_2) + (\Psi(\tau_1, \cdot), \Psi(\tau_2, \cdot))_{H^0},$$

$$E[Z(t)] = \sum_{j=1}^{q} \bar{\beta}_j A\phi_j(t),$$

$$\text{Cov}[Z(t_1), Z(t_2)] = \sum_{i,j=1}^{q} A\phi_i(t_1)Q_{ij}A\phi_j(t_2) + (A\Psi(t_1, \cdot), A\Psi(t_2, \cdot))_{H^0},$$

$$E[Y(t)] = \sum_{j=1}^{q} \bar{\beta}_j A\phi_j(t),$$

$$\text{Cov}[Y(t_1), Y(t_2)] = \sum_{i,j=1}^{q} A\phi_j(t_1)Q_{ij}A\phi_i(t_2) + (A\Psi(t_1, \cdot), A\Psi(t_2, \cdot))_{H^0}$$

$$+ R(t_1, t_2).$$
For ease of exposition let us assume that all stochastic processes are jointly normal so that we need not distinguish between wide sense conditional expectation and strict sense conditional expectations.

When $\beta$ is Bayes, the statistical inference problem is to find the conditional distribution of $\beta$ and $U(\cdot) = \{(h, U)_{H_0}, h \in H_0^c\}$, given the observed series $\{Y(t), t \in T\}$. We denote by $\beta^+$ and $Q^+$ the conditional mean and conditional covariance of $\beta$;

$$(7) \quad \beta^+ = E[\beta|Y(t), t \in T], \quad Q^+ = E[(\beta - \beta^+)(\beta - \beta^+)]Y(t), t \in T].$$

We denote by $m^+_U$ and $Q^+_U$ the conditional mean value function and conditional covariance kernel of $U(\cdot)$;

$$(8) \quad m^+_U(h) = E[(h, U)_{H_0}|Y(t), t \in T], \quad Q^+_U(h_1, h_2) = \text{Cov}[(h_1, U)_{H_0}, (h_2, U)_{H_0}|Y(t), t \in T].$$

It should be noted that the unconditional mean value function and covariance kernel of $U(\cdot)$ is

$$(9) \quad m_U(h) = E[(h, U)_{H_0}] = 0, \quad Q_U(h_1, h_2) = \text{Cov}[(h_1, U)_{H_0}, (h_2, U)_{H_0}] = (h_1, h_2)_{H_0}.$$  

When $\beta$ is non-Bayes, we choose as our estimation criterion minimum variance unbiased linear estimation of $\beta$ and minimum variance unbiased linear prediction of $U(\cdot)$. We briefly define these concepts.

An estimator $\hat{\beta}^*$ of the parameter vector $\beta$ is called unbiased if it is a function only of the observations $Y(\cdot)$ and

$$(10) \quad E[\beta|\beta^*] = \beta \quad \text{for all } \beta \in E_q,$$  

where $E_q$ is $q$-dimensional Euclidean space. $\beta^*$ is called a linear estimator if each component is a random variable belonging to $H_Y$. The minimum variance unbiased linear estimator, denoted $\hat{\beta}$, is defined to have minimum covariance matrix among all unbiased linear estimators; in symbols, for all $\beta^*$ satisfying (10)

$$(11) \quad E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top] \ll E[(\beta^* - \beta)(\beta^* - \beta)]$$

A predictor $\hat{\theta}^*$ of the random variable $\theta = X(\tau)$ is called an unbiased linear predictor if it belongs to $H_{\theta}$ and

$$(12) \quad E[\hat{\theta}^*] = E[p|\theta] \quad \text{for all } \theta \in E_q.$$  

The minimum variance unbiased linear predictor is the unbiased linear predictor, denoted $\hat{\theta}$, satisfying

$$(13) \quad E[\hat{\theta}^*|\theta^2] = \min_{\theta^*} E[\theta^*|\theta^2]$$

where the minimum is taken over all unbiased linear predictors $\theta^*$ of $\theta$. 

Associated with the minimum variance unbiased estimator $\hat{\beta}$ of $\beta$ is its covariance matrix,

$$\hat{Q} = E_p[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\prime].$$

Associated with the minimum variance unbiased predictors $\{\hat{X}(\tau), \tau \in T_X\}$ of $\{X(\tau), \tau \in T_X\}$ is the covariance kernel

$$\hat{Q}_X(\tau_1, \tau_2) = E_p[\{\hat{X}(\tau_1) - X(\tau_1)\} \{\hat{X}(\tau_2) - X(\tau_2)\}].$$

Note that $Q_X(\tau_1, \tau_2) = \text{Cov}_p[X(\tau_1), X(\tau_2)].$

The non-Bayesian solution to our statistical inference problem consists in giving formulas for $\hat{\beta}, \hat{Q}, \hat{\Sigma}, \hat{Q}_X.$

In discussing the solution to the foregoing problems, one can proceed in three stages:

1. Obtain predictors of a signal process $X$ from an observation process $Y$ under the assumption that each has known mean, covariance, and cross-covariance.

2. Under the assumption that means are known linear functions of an unknown vector parameter $\beta,$ obtain estimators of $\beta.$

3. Using the estimators of $\beta$ [given by step (2)] as if they were the known true values, obtain predictors of $X$ [using step (1)], after establishing the validity of this procedure.

In order not to overload the present paper, we will only discuss step (1); the discussion of steps (2) and (3) will be presented elsewhere.

6. RKHS solution of the zero mean linear statistical inference problem.

In studying the solution to the linear statistical inference problem it is useful to begin by omitting $\beta$ from the problem, by setting $\beta = 0.$ We thus consider an observed process

$$(1) \quad Y(t) = (\Phi(t, \cdot), U)_{-H^0} + N(t)$$

where $U$ is a $H^0$-valued white noise and $\{\Phi(t, \cdot), t \in T\}$ is a family of functions in $H^0.$ Note that the function previously denoted $A\Psi(t, \cdot)$ is now denoted $\Phi(t, \cdot).$

Fix $h \in H^0;$ the best (minimum mean square error linear) predictor of $U(h) = (h, U)_{-H^0},$ given $\{Y(t), t \in T\},$ is the random variable $m_U^+(h)$ defined as follows: $m_U^+(h)$ is the unique random variable in $H_Y = H(Y(t), t \in T)$ satisfying

$$(2) \quad E[m_U^+(h)Y(t)] = E[(h, U)_{-H^0}Y(t)], \quad t \in T.$$

But

$$(3) \quad E[(h, U)_{-H^0}Y(t)] = (\Phi(t, \cdot), h)_{H^0}.$$
Define a mapping $\Phi$ from $H^o$ to functions of $t$ by

\begin{equation}
\Phi h(t) = (\Phi(t, \cdot), h)_{H^0}.
\end{equation}

Note that (3) implies $\Phi h \in H(R + K)$ for all $h \in H^0$. We will find other Hilbert spaces to which $\Phi h$ belongs for all $h \in H^0$.

The covariance kernel of the observed process $Y(\cdot)$ is $K(t_1, t_2) + R(t_1, t_2)$; defining

\begin{equation}
K(t_1, t_2) = (\Phi(t_1, \cdot), \Phi(t_2, \cdot))_{H^0},
\end{equation}

note that $K$ is the covariance kernel of the transformed signal process $Z(t) = (\Phi(t, \cdot), U)_{-H^0}$.

The notation is now at hand for expressing the conditional mean and covariance of $U(\cdot)$ given $Y(\cdot)$:

\begin{align*}
&\mu_U^*(h) = (\Phi h, Y)_{-H(R + K)}, \\
&Q_U^*(h_1, h_2) = Q_U(h_1, h_2) - (\Phi h_1, \Phi h_2)_{H(R + K)}.
\end{align*}

To apply the general results (6) and (7) note that for a signal process $X(\cdot)$ defined by

\begin{equation}
X(t) = (\Psi(t, \cdot), U)_{H^0}
\end{equation}

the best predictor is

\begin{equation}
\hat{X}(t) = (\Phi \Psi(t, \cdot), Y)_{H(R + K)}
\end{equation}

with conditional covariance kernel

\begin{equation}
\begin{aligned}
\text{Cov}[X(t_1), X(t_2)|Y(t), t \in T] &= (\Psi(t_1, \cdot), \Psi(t_2, \cdot))_{H^0} - (\Phi \Psi(t_1, \cdot), \Phi \Psi(t_2, \cdot))_{H(R + K)}.
\end{aligned}
\end{equation}

The general results (6) and (7) are expressed in terms of the RKHS corresponding to the covariance kernel $R + K$. The inner product in this space is not usually known to us; for computation it is convenient to express $\mu_U^*$ and $Q_U^*$ in terms of $H(R)$. For this purpose one needs to make a basic assumption on the covariance kernel $K$:

\begin{equation}
K \in H(R) \otimes H(R).
\end{equation}

In words, (11) states that $K$ belongs to the RKHS which is the direct product of $H(R)$ and $H(R)$, and is denoted $H(R) \otimes H(R)$.

Direct product RKHS play a basic role in the application of RKHS methods to time series analysis: I believe their first application was by myself [(1963), p. 164, reprinted in (1967), p. 486] to provide a simple formulation of conditions for the equivalence of two normal processes with unequal covariance kernels [see § 11]. For the purposes of the present
exposition we use the following definition of the direct product $H(R_1) \otimes H(R_2)$ of the RKHS corresponding to two covariance kernels $R_1$ and $R_2$, since it is the most convenient for proving theorems. Let $H(R_i)$ consist of functions on an index set $T_i$,

\[
\{ \phi_x \} \text{ be any CONS in } H(R_1), \\
\{ \psi_\beta \} \text{ be any CONS in } H(R_2),
\]

where CONS denotes complete orthonormal set of functions; then

\[
H(R_1) \otimes H(R_2) = \left\{ h \text{ on } T_1 \otimes T_2 : h(t_1, t_2) = \sum_{x, \beta} h_{x, \beta} \phi_x(t_1) \psi_\beta(t_1) \right\}
\]

for some double sequence $\{ h_{x, \beta} \}$ such that $\sum_{x, \beta} h_{x, \beta}^2 < \infty$.

The norm of a function $h$ in $H(R_1) \times H(R_2)$ is given by

\[
\| h \|_{H(R_1) \otimes H(R_2)}^2 = \sum_{x, \beta} h_{x, \beta}^2.
\]

A kernel $K$ in $H(R) \otimes H(R)$ generates an operator $K : H(R) \to H(R)$ by, for every $f \in H(R)$,

\[
(12) \quad Kf(t) = (K(\cdot, t), f)_{H(R)}.
\]

The operator $K$ is Hilbert Schmidt (for a definition of the notion of Hilbert Schmidt operator see any text on functional analysis or integral equations; for example, Riesz-Nagy, (1955), p. 242).

It is of interest to relate the operator $K$ to the operator $\Phi$ defined by (4) when

\[
(13) \quad K(s, t) = (\Phi(s, \cdot), \Phi(t, \cdot))_{H^0}.
\]

Therefore (12) can be written (formally)

\[
Kf(t) = (K(s, t), f(s))_{H(R)} = ((\Phi(s, u), \Phi(t, u))_{H^0}, f(s))_{H(R)}
\]

\[
(14) \quad = ((\Phi(s, u), f(s))_{H(R)}, \Phi(t, u))_{H^0} = \Phi \Phi^* f(t)
\]

defining $\Phi^* : H(R) \to H^0$ by

\[
(15) \quad \Phi^* f(u) = (\Phi(s, u), f(s))_{H(R)}.
\]

The notation $\Phi^*$ is meant to indicate that $\Phi^*$ is the adjoint of $\Phi$; see (20).

To make the foregoing assumptions rigorous we make the basic assumption that the functions $\{ \Phi(t, \cdot), t \in T \}$ are selected as functions of two variables $t$ and $u$ so that

\[
(16) \quad \Phi(t, u) \in H(R) \otimes H^0
\]
or equivalently that
\begin{equation}
\Phi(t, u) = \sum_{\alpha, \beta} \Phi_{\alpha\beta} \phi_\alpha(t) \psi_\beta(u)
\end{equation}

for some CONS \{\phi_\alpha\} in \(H(R)\), CONS \{\psi_\beta\} in \(H^0\), and double sequence \{\Phi_{\alpha\beta}\} satisfying
\begin{equation}
\sum_{\alpha, \beta} \Phi_{\alpha\beta}^2 < \infty;
\end{equation}

then (4) defines a Hilbert Schmidt operator
\begin{equation}
\Phi: H^0 \to H(R)
\end{equation}

with adjoint \(\Phi^*: H(R) \to H^0\) given by (15), in the sense that for every \(h \in H^0\) and \(f \in H(R)\)
\begin{equation}
(\Phi h, f)_{H(R)} = (h, \Phi^* f)_{H^0}.
\end{equation}

We briefly outline the proof of these assertions. Write
\begin{equation}
h(u) = \sum_\beta h_\beta \psi_\beta(u), \quad f(t) = \sum_\alpha f_\alpha \phi_\alpha(t).
\end{equation}

Then
\begin{equation}
\Phi h(t) = \sum_\beta h_\beta \sum_\alpha \Phi_{\alpha\beta} \phi_\alpha(t) = \sum_\alpha \phi_\alpha(t) \sum_\beta \Phi_{\alpha\beta} h_\beta
\end{equation}

and
\begin{equation}
(\Phi h, f) = \sum_{\alpha, \beta} f_\alpha \Phi_{\alpha\beta} h_\beta;
\end{equation}

similarly
\begin{equation}
\Phi^* f(u) = \sum_\alpha f_\alpha \sum_\beta \Phi_{\alpha\beta} \psi_\beta(u) = \sum_\beta \psi_\beta(u) \sum_\alpha f_\alpha \Phi_{\alpha\beta}
\end{equation}

and
\begin{equation}
(h, \Phi^* f)_{H^0} = \sum_{\alpha, \beta} f_\alpha \Phi_{\alpha\beta} h_\beta.
\end{equation}

We next prove that (16) implies that
\begin{equation}
K \in H(R) \otimes H(R)
\end{equation}

and
\begin{equation}
K = \Phi \Phi^*.
\end{equation}
To prove (26) and (27) we use (13) and (17) to express $K(s, t)$ in a double series in the CONS $\{\phi_i\}$:

$$K(s, t) = \sum_{\beta} \sum_{\xi_1} \Phi_{\xi_1 \beta} \phi_{\xi_1}(s) \sum_{\xi_2} \Phi_{\xi_2 \beta} \phi_{\xi_2}(t) = \sum_{\xi_1 \xi_2} K_{\xi_1 \xi_2} \phi_{\xi_1}(s) \phi_{\xi_2}(t)$$

defining

$$K_{\xi_1 \xi_2} = \sum_{\beta} \Phi_{\xi_1 \beta} \Phi_{\xi_2 \beta}.$$

The coefficients $\{K_{\xi_1 \xi_2}\}$ satisfy

$$\sum_{\xi_1 \xi_2} |K_{\xi_1 \xi_2}|^2 < \infty, \quad \sum_{\xi} |K_{\xi}| < \infty$$

since

$$|K_{\xi_1 \xi_2}|^2 \leq \sum_{\beta_1} \Phi_{\xi_1 \beta_1}^2 \sum_{\beta_2} \Phi_{\xi_2 \beta_2}^2, \quad |K_{\xi}| \leq \sum_{\beta} \Phi_{\xi \beta}^2.$$

From (28) and (30) we infer (26). To prove (27), note that

$$Kf(t) = \sum_{\xi_1} f_{\xi_1} \sum_{\xi_2} K_{\xi_1 \xi_2} \phi_{\xi_2}(t)$$

while

$$\Phi \Phi^* f(t) = \sum_{\xi_2} \phi_{\xi_2}(t) \sum_{\beta} \phi_{\xi_2 \beta} \sum_{\xi_1} f_{\xi_1} \Phi_{\xi_1 \beta}.$$

So far we have merely defined the operator $K$ and related it to $\Phi$. Next let us state the basic relation between $H(R + K)$ and $H(R)$ that follows from (26): $H(R + K)$ consists of the same functions as $H(R)$ and for every $f$ and $g$ in $H(R)$

$$(f, g)_{H(R + K)} = ((I + K)^{-1} f, g)_{H(R)}.$$

We outline a proof of (34); let $f_1 = (I + K)^{-1} f$ so that $(I + K)f_1 = f$. One sees that (34) is equivalent to

$$(f_1, g)_{H(R + K)} = (f_1, g)_{H(R)}$$

for all $f_1$ and $g$ in $H(R)$. To prove (35) it suffices to prove it for $g = R + K(\cdot, t)$; with this choice of $g$, both sides of (35) equal $f_1(t) + Kf_1(t) = f(t)$. To summarize the argument, we have shown that the inner product defined on the functions belonging to $H(R + K)$ by the right-hand side of (34) is the correct inner product since it has the reproducing property

$$(f, R + K(\cdot, t))_{H(R + K)} = f(t).$$
The facts are at hand to conclude our discussion of the zero mean linear statistical inference problem. Assuming (16), the conditional mean and covariance of $U(\cdot)$ given $Y(\cdot)$ is given by

$$m_U^+(h) = ((I + \Phi\Phi^*)^{-1}\Phi h, Y)_{H(R)},$$

$$Q_U(h_1, h_2) = Q_U(h_1, h_2) - ((I + \Phi\Phi^*)^{-1}\Phi h_1, \Phi h_2)_{H(R)}.$$  

A comment is necessary on the meaning of $(f, Y)_{H(R)}$ for any $f \in H(R)$. For $f$ a simple function in the sense that (for some integer $n$, reals $c_i$, and indices $t_i$)

$$f = \sum_{i=1}^{n} c_i R(\cdot, t_i)$$

we define

$$(f, Y)_{H(R)} = \sum_{i=1}^{n} c_i Y(t_i).$$

For $f$ a limit in $H(R)$-norm of simple functions $f_n$, we define

$$ (f, Y)_{H(R)} = \lim_{n} (f_n, Y)_{H(R)} $$

where this limit is in the Hilbert space of random variables with inner product corresponding to the covariance kernel $R + K$. For $f$ in (39) and random variable in (40), one can verify that

$$E[(f, Y)_{H(R)}]^2 = \sum_{i,j} c_i c_j \{R(t_i, t_j) + K(t_i, t_j)\} = ((I + K)f, f)_{H(R)}.$$  

One can show that $\{(f, Y)_{H(R)}, f \in H(R)\}$ is a Hilbert space indexed collection of random variables with mean zero and covariance operator $I + K$.

Finally, we turn to an important interpretation of (37) which will enable us to show the equivalence of the solution of various approximation and control problems. In this study a central role will be played by the operator identity

$$(I + \Phi\Phi^*)^{-1}\Phi = \Phi(I + \Phi^*\Phi)^{-1};$$

to verify this identity multiply on the left by $I + \Phi\Phi^*$ and multiply on the right by $I + \Phi^*\Phi$.

Using (43) rewrite the formula (37) for $m_U^+$:

$$m_U^+(h) = (\Phi(I + \Phi^*\Phi)^{-1}h, Y)_{H(R)}.$$  

To take the next step we must act as if the inner product in (44) is a real
inner product in $H(R)$, rather than a congruence inner product. Then we can express it as an inner product in $H^0$:

\begin{equation}
    m_U^+(h) = (h, (I + \Phi^*\Phi)^{-1}\Phi^*Y)_{H^0}.
\end{equation}

Our formula for $Q_U^+$ can be rigorously rewritten in terms of inner products in $H^0$:

\begin{equation}
    Q_U^+(h_1, h_2) = (h_1, h_2)_{H^0} - (\Phi^*\Phi(I + \Phi^*\Phi)^{-1}h_1, h_2)_{H^0} \\
    = ((I + \Phi^*\Phi)^{-1}h_1, h_2)_{H^0}.
\end{equation}

The parameter process $U(\cdot)$ regarded as a random element with values in $H^0$ thus has conditional mean

\begin{equation}
    \hat{U} = (I + \Phi^*\Phi)^{-1}\Phi^*Y
\end{equation}

and a conditional covariance operator equal to $(I + \Phi^*\Phi)^{-1}$. In the next section we show that $\hat{U}$, which at this stage has no rigorous meaning since the sample path of the process $Y$ never belongs to $H(R)$, is the solution to a suitable optimization problem.

7. Equivalence between time series parameter estimation, control theory, and approximation theory. There is a remarkable equivalence between the solutions to the problems of regression analysis, approximation by spline functions, control of linear plants with quadratic cost criterion, Kalman filtering, solution of integral equations, and time series parameter estimation.

The first observation that leads to this equivalence is to note that all the foregoing problems (in their simplest versions) are of the form: “solve” for $U$ in the linear system

\begin{equation}
    Y = \Phi U
\end{equation}

where $Y$ is the observation or measurement (and is therefore considered a known function) and $\Phi$ is a known linear operator. To make this problem more precise we must specify the domain and range spaces of the operator $\Phi$. We find it clearer to leave open the range space of $\Phi$ and to specify it as a collection $\{\Phi(t), t \in T\}$ of elements in a Hilbert space $H^0$. Then we seek to find that $U$ in $H^0$ satisfying

\begin{equation}
    Y(t) = (\Phi(t), U)_{H^0}, \quad t \in T.
\end{equation}

We continue to write the linear system (2) intuitively as $Y = \Phi U$.

There are various senses in which one can solve the linear system (2) for $U$; our basic classification is in terms of

\begin{equation}
    \text{interpolation vs. smoothing,}
\end{equation}

\begin{equation}
    \text{non-Bayes unknown vs. Bayes unknown.}
\end{equation}
These names are a blend of ideas from the theory of approximation (of a function given its values at a set of points) and the theory of statistical inference. It is beyond the scope of this paper to define these terms; we wish merely to indicate the solutions to \( Y = \Phi U \) to which they lead:

<table>
<thead>
<tr>
<th>Interpolation</th>
<th>Smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Bayes ( U )</td>
<td>( U = \Phi^{-1} Y )</td>
</tr>
<tr>
<td></td>
<td>( U = (\Phi^* \Phi)^{-1} \Phi^* Y )</td>
</tr>
<tr>
<td>Bayes ( U )</td>
<td>( U = \Phi^<em>(\Phi \Phi^</em>)^{-1} Y )</td>
</tr>
<tr>
<td></td>
<td>( U = (\lambda I + \Phi^* \Phi)^{-1} \Phi^* Y )</td>
</tr>
<tr>
<td></td>
<td>( = \Phi^<em>(\lambda I + \Phi \Phi^</em>)^{-1} Y )</td>
</tr>
</tbody>
</table>

The problem we desire to consider in detail is the optimization problem: given a Hilbert space \( H^0 \), and index set \( T \), a family of elements \( \{ \Phi(t), t \in T \} \) in \( H^0 \), and RKHS \( H(R) \) consisting of functions on \( T \), with reproducing kernel \( R \), and a function \( Y \in H(R) \), find \( U \) in \( H^0 \) to minimize (for a fixed positive constant \( \lambda \))

\[
J(U) = \| Y - \Phi U \|_{H(R)}^2 + \lambda \| U \|_{H^0}^2.
\]

In order to guarantee that

\[
(5) \quad \Phi U(t) = (\Phi(t), U)_{H^0}
\]

is a function in \( H(R) \) we assume that \( \Phi \) belongs to \( H(R) \otimes H^0 \), regarded as a function of two variables. Then \( \Phi : H^0 \to H(R) \) is a Hilbert Schmidt operator with adjoint \( \Phi^* : H(R) \to H^0 \). To minimize \( J(U) \) one writes it

\[
J(U) = \| Y \|_{H(R)}^2 - 2\langle Y, \Phi U \rangle_{H(R)} + (\Phi U, \Phi U)_{H(R)} + \lambda \langle U, U \rangle_{H^0}
\]

\[
= \| Y \|_{H(R)}^2 - 2\langle \Phi^* Y, U \rangle_{H^0} + ((\lambda I + \Phi^* \Phi)U, U)_{H^0}
\]

which is minimized by

\[
(7) \quad \hat{U} = (\lambda I + \Phi^* \Phi)^{-1} \Phi^* Y
\]

with

\[
J(\hat{U}) = \| Y \|_{H(R)}^2 - ((\lambda I + \Phi^* \Phi)\hat{U}, \hat{U})_{H^0}
\]

\[
= \| Y \|_{H(R)}^2 - (\Phi Y, \Phi(\lambda I + \Phi^* \Phi)^{-1} \Phi^* Y)_{H(R)}
\]

\[
= \| Y \|_{H(R)}^2 - (Y, (\lambda I + \Phi \Phi^*)^{-1} \Phi^* Y)_{H(R)}
\]

\[
= \lambda ((\lambda I + \Phi \Phi^*)^{-1} Y, Y)_{H(R)}.
\]

Comparing equation (47) of the previous section with equation (7) of this section one sees that the two answers are equivalent. It is widely accepted among engineers concerned with filtering theory that the solution to the time series parameter estimation can be found by solving a
suitable optimization problem of the form (4); see for example Bensoussan (1969). However, no rigorous proof of this equivalence has been given in general. I believe that a rigorous proof can be given using the ideas we have discussed in this paper.

Chapter 3. Examples of RKHS

8. Integral representation theorem. The basic approach to finding the RKHS associated with a covariance kernel $K$ on $T \otimes T$ is to find an integral representation for $K$ of the form

$$K(s, t) = \int_{\lambda} g(s, \lambda) g(t, \lambda) \mu(d\lambda)$$

(1)

where $\mu$ is a measure and $\{g(t, \cdot), t \in T\}$ is a family of functions in $L_2(\mu) = \{\text{measurable } f \text{ with domain } \Lambda: \int_{\Lambda} f^2(\lambda) d\lambda < \infty\}$. When (1) holds, $H(K)$ consists of all functions $f$ on $T$ of the form

$$f(t) = \int_{\Lambda} F(\lambda) g(t, \lambda) \mu(d\lambda)$$

(2)

for some unique $F$ in $H_2 = H(g(t, \cdot), t \in T)$, the Hilbert subspace of $L_2(\mu)$ spanned by $\{g(t, \cdot), t \in T\}$. The RKHS norm of $f$ is given by

$$\|f\|_{H(K)}^2 = \|F\|_{L_2(\mu)}^2.$$  

(3)

To prove (2) we merely note that

$$K(s, t) = (K(\cdot, s), K(\cdot, t))_{H(K)} = (g(s, \cdot), g(t, \cdot))_{L_2(\mu)}$$

implies that there is a congruence (one-to-one inner product preserving mapping) between $H(K)$ and $H_2$ such that $K(\cdot, t)$ and $g(t, \cdot)$ transform into each other. If $f \in H(K)$ and $F \in L_2(\mu)$ are transforms of each other,

$$f(t) = (f, K(\cdot, t))_{H(K)} = (F, g(t, \cdot))_{L_2(\mu)}$$

(4)

which proves (2).

To illustrate the use of the integral representation approach, let us consider the Wiener process $\{W(t), 0 \leq t < T\}$ which has covariance kernel

$$K_W(s, t) = \min(s, t).$$

(6)

This kernel has integral representation

$$K(s, t) = \int_{0}^{T} (s - u)^+_0 (t - u)^+_u du$$

(7)
where for $k \geq 0$

\[ x_+^k = 0 \text{ if } x \leq 0, \]
\[ = x^k \text{ if } x > 0 \]

so that $(t - u)_+^0 = 1$ or $0$ according as $u < t$ or $u \geq t$.

From (7) it follows that $H(K)$ consists of functions $f$ of the form

\[ f(t) = \int_0^T F(u)(t - u)_+^0 \, du = \int_0^T F(u) \, du, \quad 0 \leq t \leq T, \]

which are indefinite integrals of functions $F$ which are square integrable on the interval $0 \leq t \leq T$; the norm of $f$ is

\[ \| f \|_{H(K)}^2 = \int_0^T |F(u)|^2 \, du. \]

Formally $F(t) = f'(t)$.

Next consider a process $\{X(t), 0 \leq t \leq 1\}$ with covariance kernel

\[ K(s, t) = \int_0^1 (q!)^{-1} (s - \lambda)_+^q (t - \lambda)_+^q \, d\lambda \]

where $q$ is a fixed integer. It follows that $H(K)$ consists of all functions $f$ of the form

\[ f(t) = \int_0^1 F(\lambda) \frac{1}{q!} (t - \lambda)_+^q \, d\lambda \]

for some square integrable function $F$. Formally

\[ f^{(q+1)}(t) = F(t) \]

and $H(K)$ consists of functions $f$ whose $(q + 1)^{st}$ derivative is square integrable. The process $X(\cdot)$ can be represented in terms of the Wiener process $W(\cdot)$ by

\[ X(t) = \int_0^1 \frac{1}{(q - 1)!} (t - \lambda)_+^{q-1} W(\lambda) \, d\lambda; \]

in terms of white noise $W(\lambda)$, the formal derivative of $W(\lambda)$. One can represent $X(t)$ by

\[ X(t) = \int_0^1 \frac{1}{q!} (t - \lambda)_+^q W(\lambda) \, du. \]
The process $X(\cdot)$ defined by (15) satisfies the formal stochastic differential equation

$$
X^{(q)}(t) = \tilde{W}(t);
$$

in words, the $q$th derivative of $X(\cdot)$ is white noise. The most general solution of (16) is

$$
X(t) = \beta_1 + \beta_2 + \ldots + \beta_q t^{q-1} + \int_0^t \frac{1}{\gamma^q} (t - \lambda)^{q-1} \tilde{W}(\lambda) d\lambda.
$$

This is an example of a process $X(\cdot)$ with representation (5.1).

9. Hilbert space indexed time series. Let $H^0$ be a separable Hilbert space, and $\{U(h), h \in H^0\}$ an $H^0$-indexed series with zero mean and covariance operator $Q$:

$$
Q_U(h_1, h_2) = \text{Cov}[U(h_1), U(h_2)] = (Qh_1, h_2)_{H^0}.
$$

Let $\Phi = \Phi^* = Q^{1/2}$ so that $Q = \Phi \Phi^*$. Then

$$
Q_U(h_1, h_2) = (\Phi h_1, \Phi h_2)_{H^0}.
$$

Define

$$
H_\Phi = \overline{\Phi H^0} \equiv \text{closure of } \{g : g = \Phi h \text{ for some } h \in H^0\}.
$$

The RKHS corresponding to $Q$ is

$$
H(Q) = \{f \text{ on } H^0 : f(h) = \langle \Phi h, F \rangle_{H^0} \text{ for some } f \in H_\Phi, \|F\|_{H^0}^2 = \|F\|_{H_\Phi}^2\}.
$$

In words, the RKHS of an $H^0$-indexed series is a family of linear functionals, so that $H(Q)$ is set theoretically a subset of $H^0^*$, the topological dual space of $H^0$ (which is congruent to $H^0$). There is a subset of $H^0$, denoted $H_\sim(Q)$, with which $H(Q)$ is congruent:

$$
H_\sim(Q) = \{g \in H^0 : g = \Phi^* F \text{ for some } F \in H_\Phi, \|F\|_{H_\sim(Q)}^2 = \|F\|_{H^0}^2\}.
$$

We call $H_\sim(Q)$ a congruent RKHS, with reproducing kernel $Q$. Note that a RKHS is always a space of functions, while a congruent RKHS is a subset of a Hilbert space $H^0$ which is congruent to a RKHS of functions on $H^0$.

One can characterize $H_\sim(Q)$ more concretely in the case that $\Phi$ is a Hilbert Schmidt operator; then one can find an orthonormal set $\{\phi_j\}$ in $H^0$ and a sequence of nonnegative constants $\lambda_j$ such that $\phi_j$ and $\lambda_j$ are respectively eigenfunctions and eigenvalues of $Q$:

$$
Q\phi_j = \lambda_j \phi_j, \quad \Phi \phi_j = \lambda_j^{1/2} \phi_j.
$$
Note that for every $h \in H^0$

$$h = \sum (\phi_j, h)_{H^0} \phi_j, \quad \Phi h = \sum \lambda_j^{1/2} (\phi_j, h)_{H^0} \phi_j. \quad (7)$$

Assume further that the range of $\Phi$ is dense in $H^0$ so that $H_\Phi = H^0$ (this assumption is made only for ease of exposition and can be dispensed with). Then $H_\sim(Q)$ consists of

$$\{ g \in H^0 : g = \sum \lambda_j^{1/2} (\phi_j, F)_{H^0} \phi_j \text{ for some } f \in H^0, \}$$

$$\| g \|_{H_\sim(Q)}^2 = \sum \| (\phi_j, F)_{H^0} \|^2. \quad (8)$$

But one can write $g$ in terms of its own Fourier coefficients:

$$g = \sum g_j \phi_j, \quad g_j = (\phi_j, g)_{H^0}. \quad (9)$$

One concludes that

$$H_\sim(Q) = \left\{ g \in H^0 : g = \sum g_j \phi_j \text{ for some sequence } \{g_j\} \text{ such that} \right\}$$

$$\| g \|_{H_\sim(Q)}^2 = \sum \frac{1}{\lambda_j} g_j^2 < \infty \right\}. \quad (10)$$

It should be noted that every Hilbert space $H^0$ is the congruent RKHS of $H^0$-indexed white noise; when $Q = I$ we obtain from (5)

$$H_\sim(I) = \{ g \in H^0 : g = F \text{ for some } f \in H^0, \| g \|_{H_\sim(I)}^2 = \| F \|_{H^0}^2 \}. \quad (11)$$

$L_2$-spaces (Hilbert spaces of square integrable functions on some measure space) differ fundamentally from RKHS in the fact that in $L_2$-spaces convergence in norm of a sequence of functions does not imply pointwise convergence of the sequence. The celebrated Dirac delta function $\delta(\cdot)$ is the "reproducing kernel" of the $L_2$-space of square integrable functions on the real line; symbolically one can consider white noise to have covariance kernel $\delta(t - s)$.

Finally, it should be noted that a Hilbert space whose members are functions on a set $T$, which is a congruent RKHS in the sense of (10), can be the true RKHS of a time series $\{ Y(t), t \in T \}$ with index set $T$. For example, corresponding to $\{ Y(t), 0 \leq t \leq 1 \}$ with zero means and continuous kernel $Q(s, t)$, we can define an operator $Q$ on $H^0 \equiv L_2[0, 1]$ by

$$Qf(t) = \int_0^1 Q(s, t) f(s) ds, \quad 0 \leq t \leq 1. \quad (12)$$

Let $\{ \phi_j \}$ and $\{ \lambda_j \}$ be the eigenfunctions and positive eigenvalues of the
integral equation

\begin{equation}
Q\phi_j(t) = \int_0^1 Q(s,t)\phi_j(s)\, ds = \lambda_j\phi_j(t), \quad 0 \leq t \leq 1.
\end{equation}

Assume further that \( Q \) is positive definite in the sense that for any function \( f \) in \( H^0 \)

\begin{equation}
\int_0^1 \int_0^1 Q(s,t)f(s)f(t)\, ds\, dt = 0 \quad \text{implies } f(t) = 0 \text{ a.e.}
\end{equation}

Then the RKHS corresponding to the kernel \( Q(s,t) \) is

\begin{equation}
H(Q) = \left\{ g \in H^0 : \|g\|_{H(Q)}^2 = \sum \frac{1}{\lambda_j} \left| \int_0^1 \phi_j(t) g(t)\, dt \right|^2 < \infty \right\}
\end{equation}

which is the right-hand side of (10).

10. Normal Markov processes. A normal process \( \{Y(t), 0 \leq t \leq T\} \) with zero means and continuous covariance kernel \( K(s,t) \) can be shown to be Markov if and only if

\begin{equation}
K(s,t) = g(s)G(\min(s,t))g(t)
\end{equation}

where

(i) \( g \) is nonvanishing, continuous, and \( g(0) = 1 \), and
(ii) \( G \) is continuous and monotone increasing.

Important examples of normal Markov processes are:

- Wiener: \( K(s,t) = \sigma^2 \min(s,t) \), \( g(t) = 1 \), \( G(t) = \sigma^2 t \);
- Stationary Markov: \( K(s,t) = \sigma_0^2 e^{-\beta|t-s|} \), \( g(t) = e^{-\beta t} \), \( G(t) = \sigma_0^2 e^{2\beta t} \);
- Pinned Wiener (\( T = 1 \)): \( K(s,t) = \min(s,t) - st \), \( g(t) = 1 - t \), \( G(t) = t/(1 - t) \).

A covariance kernel of the form (1) has the following (Stieltjes) integral representation:

\begin{equation}
K(s,t) = \int_0^T g(s)(s-u)^0 g(t)(t-u)^0 \, dG(u) + g(s)g(t)G(0).
\end{equation}

Let us assume \( G(0) > 0 \) since this is slightly more complicated than the case \( G(0) = 0 \).

From the integral representation theorem, it follows that \( H(K) \) consists of all functions \( f \) of the form

\begin{equation}
f(t) = \int_0^T g(t)(t-u)^0 F(u) \, dG(u) + g(t)F(0)G(0)
\end{equation}
for some constant $F_0$ and function $F(u)$ in $\{h: \int_0^T |h(u)|^2 \, dG(u) < \infty\}$; the RKHS norm of $f(\cdot)$ is

\[
\|f\|^2_{H(K)} = \int_0^T |F(u)|^2 \, dG(u) + F_0^2 G(0).
\]

To be more concrete, let us assume that $g$ and $G$ are differentiable; then (3) yields

\[
G'(t)F(t) = \left\{ \frac{f(t)}{g(t)} \right\}', \quad F_0 = \frac{f(0)}{g(0)G(0)} = \frac{f(0)}{G(0)}
\]

since $g(0) = 1$. One can now show that $H(K)$ consists of all $L_2$-differentiable functions $f(\cdot)$ with domain $0 \leq t \leq T$, such that

\[
\|f\|^2_{H(K)} = \int_0^T \left\{ \frac{f(t)}{g(t)} \right\}^2 \frac{1}{G'(t)} \, dt + |f(0)|^2 \frac{1}{G(0)} < \infty.
\]

Formula (6) has been given by several authors; one important reference is Sacks and Ylvisaker (1966), p. 86.

In my view a more physically meaningful formula for $\|f\|^2_{H(K)}$ is found by introducing

\[
\rho(t) = g'(t)/g(t), \quad \sigma^2(t) = g^2(t)G'(t)
\]

in terms of which we can write the $H(K)$ inner product

\[
(f_1, f_2)_{H(K)} = \int_0^T \left\{ f_1'(t) - \rho(t)f_1(t) \sigma^{-2}(t) \right\} \left\{ f_2'(t) - \rho(t)f_2(t) \right\} \, dt
\]

\[+ f_1(0)G^{-1}(0)f_2(0).\]

When one considers problems of statistical inference for normal Markov processes, it appears that $\rho(t)$ and $\sigma^2(t)$ are the natural parameters to estimate; for this reason we regard them as the natural parameters in terms of which to express the RKHS inner product.

**Chapter 4.**

**Probability Density Functionals of Normal Processes**

In this review many areas of time series analysis to which RKHS have been applied can only be indicated by reference to the literature. RKHS have been used to:

1. Characterize the support of the probability measure on a Banach space induced by a normal time series [see Kallianpur (1970a)];
2. Represent an arbitrary nonlinear functional of a normal process (Cameron-Martin expansion and multiple Wiener integral expansion) [see Neveu (1968), Kallianpur (1970b), Hida-Ikeda (1967)];

3. Represent a normal process in terms of white noise processes representing "innovations" in the process [see Hida (1960)];

4. Obtain zero-one laws for normal processes [see Kallianpur (1970c)];

5. Determine conditions for convergence in distribution of stochastic processes [see Brown (1969)];

6. Obtain statistical designs and quadrature formulas [see Sacks and Ylvisaker (1966), (1968), (1969) and Wahba (1969a)];

7. Establish equivalence of time series prediction and spline functions [see Kimeldorf and Wahba (1968, 1969)];

8. Provide algorithms for numerical solution of integral and differential equations [see Wahba (1969b)];

9. Provide coordinate free approach to statistical communication theory [see Kailath (1969)].

It was mentioned in the introduction (§ 0) that a major reason for the usefulness of RKHS is that a basic tool in many of the foregoing problem areas is the theory of equivalence and singularity of normal measures. The aim of this chapter is to indicate how the main results of this theory can be simply expressed in terms of RKHS.

11. Equivalence and orthogonality of normal processes. Given two probability measures $P_1$ and $P_0$ on a measurable space $(\Omega, \mathcal{A})$ we say that

(i) $P_1$ is absolutely continuous with respect to $P_0$, denoted $P_1 \ll P_0$, if

\[ P_1[A] = 0 \text{ implies } P_0[A] = 0. \tag{1} \]

or equivalently if there exists a measurable function $p$ such that for all $A \in \mathcal{A}$

\[ P_1[A] = \int_A p \, dP_0; \tag{2} \]

(ii) $P_1$ and $P_0$ are equivalent, denoted $P_1 \equiv P_0$, if $P_1 \ll P_0$ and $P_0 \ll P_1$;

(iii) $P_1$ and $P_0$ are orthogonal, denoted $P_1 \perp P_0$, if there exists a set $A$ in $\mathcal{A}$ such that

\[ P_1[A] = 1 \quad \text{and} \quad P_0[A] = 0. \tag{3} \]

When (1) holds, we call $p$ the Radon-Nikodým derivative of $P_1$ with respect to $P_0$, denoted $p = dP_1/dP_0$.

To apply these notions to a stochastic process $Y(\cdot)$ we first define its probability distribution.
Given a stochastic process with index set $T$, such as $\{Y(t), t \in T\}$, by the probability distribution induced by the process we mean the probability space $(\Omega_T, \mathcal{B}_T, P_Y)$, where

$\Omega_T$ is the linear space of all real valued functions on $T$,  
$\mathcal{B}_T$ is the sigma field of cylinder sets of $\Omega_T$, defined as the smallest sigma field containing all sets of the form $\{\omega \in \Omega_T : \omega(t) \leq x\}$ for any $t \in T$ and real number $x$,

$P_Y$ is a probability measure with domain $\mathcal{B}_T$ such that 

$$P_Y(\{\omega \in \Omega_T : \omega(t_1) \leq x_1, \ldots, \omega(t_n) \leq x_n\}) = P[Y(t_1) \leq x_1, \ldots, Y(t_n) \leq x_n]$$

for any integers $n$, points $t_1, \ldots, t_n$ in $T$ and real numbers $x_1, \ldots, x_n$.

When considering the relations between normal probability measures, we denote by $P_{m,R}$ the probability measure on $(\Omega_T, \mathcal{B}_T)$ which is the probability distribution of a normal time series $Y(\cdot) = \{Y(t), t \in T\}$ with mean $m$ and covariance $R$. The following notions are needed:

(i) $H(R)$ or RKHS$(R)$, the reproducing kernel Hilbert space corresponding to $R$;

(ii) $H(R) \otimes H(R)$, the direct product RKHS;

(iii) equivalence class of $R$ defined to be the set of all covariance kernels $R$, which are equivalent to $R$.

We define two covariance kernels $R_1$ and $R_2$ to be equivalent, denoted $R_1 \equiv R_2$, if there exist constants $C_1$ and $C_2$ such that 

$$C_1 \sum_{i,j=1}^{n} c_i c_j R_1(t_i, t_j) \leq \sum_{i,j=1}^{n} c_i c_j R_2(t_i, t_j)$$

$$\leq C_2 \sum_{i,j=1}^{n} c_i c_j R_1(t_i, t_j)$$

for every integer $n$, indices $t_1, \ldots, t_n$ in $T$, and reals $c_1, \ldots, c_n$.

The basic facts about probability density functionals of normal processes are expressed in the following theorems for any mean value functions $m, m_1, m_2$ and covariance kernels $R, R_1, R_2$.

**DICHOTOMY THEOREM.** Either $P_{m_1, R_1} \equiv P_{m_2, R_2}$ or $P_{m_1, R_1} \perp P_{m_2, R_2}$; see Feldman (1958) and Hajek (1958).

Equivalence of normal processes with unequal means, equal covariances. $P_{m,R} = P_{0,R}$ if and only if $m \in H(R)$; then 

$$p = \frac{dP_{m,R}}{dP_{0,R}} = \exp\{(m, Y)_{\sim H(R)} - \frac{1}{2}||m||^2_{H(R)}\}.$$
Equivalence of normal processes with equal means, unequal covariances. 
\( P_{0, R_1} \equiv P_{0, R_2} \) if and only if

\[
R_1 \equiv R_2 \quad \text{and} \quad R_2 - R_1 \in H(R_1) \otimes H(R_1);
\]

the probability density functional is of the form

\[
p = dP_{0, R_2}/dP_{0, R_1} = Ce^J
\]

where \( C \) is a constant and \( J \) is a random variable which is a quadratic form in the sense that it belongs to the Hilbert space spanned by the family of random variables \( \{ Y(s)Y(t) - R_1(s, t), (s, t) \in T \otimes T \} \). One can express \( p \) in a variety of ways; see Shepp (1965), Varberg (1966), Golosov (1966), Neveu (1968), Rozanov (1968), Kailath (1969).

It seems difficult to claim credit for a new formula for the probability density functional, but let me note without proof what I believe to be a new version of the formula for \( p \) in the case when both normal measures correspond to Markov processes on \( 0 \leq t \leq T \), so that the representation (10.1) holds:

\[
R_1(s, t) = g_1(s)G_1(\min(s, t))g_1(t),
\]

\[
R_2(s, t) = g_2(s)G_2(\min(s, t))g_2(t)
\]

where, for \( i = 1, 2, g_i(\cdot) \) is nonvanishing, continuous, and \( g_i(0) = 1 \), and \( G_i(\cdot) \) is continuous and monotone increasing. Assume \( g_i(\cdot) \) and \( G_i(\cdot) \) have continuous derivatives, and define the “natural parameters” [used in (10.8)]

\[
\rho_i(t) = g_i'(t)/g_i(t), \quad \sigma_i^2(t) = g_i^2(t)G_i'(t).
\]

A necessary and sufficient condition for equivalence of the normal measure is

\[
\sigma_1^2(t) = \sigma_2^2(t) \equiv \sigma^2(t) \quad \text{for all } t
\]

and

\[
G_1(0), G_2(0) \quad \text{both positive or both zero.}
\]

Then, in the case that \( G_1(0) \) and \( G_2(0) \) are both positive

\[
p = \frac{G_2(0)}{G_1(0)} \exp \left[ -\frac{1}{2} Y^2(0) \left\{ \frac{1}{G_2(0)} - \frac{1}{G_1(0)} \right\} \right]
\]

\[
\cdot \exp \left[ \int_0^T \sigma^{-2}(\tau)[\rho_2(\tau) - \rho_1(\tau)]Y(\tau) \, dY(\tau) \right.
\]

\[
- \frac{1}{2} \int_0^T \sigma^{-2}(\tau)[\rho_1^2(\tau) - \rho_2^2(\tau)]Y^2(\tau) \, d\tau
\]
where the first integral in the exponent is an Itô stochastic integral and the second integral is an ordinary stochastic integral.

A good summary of the diverse RKHS conditions available for equivalence and singularity of normal processes is given by Golosov-Tempelman (1969).

12. Covariance kernels of probability density functionals. A field in which many different reproducing kernel Hilbert spaces appear simultaneously in elegant interplay is the theory of unbiased minimum variance estimation of parameters of normal processes (see Parzen (1967), p. 336). This theory is now ripe for practical application because of the availability of the following basic formulas for the covariance kernels of probability density functionals.

THEOREM: UNEQUAL MEANS [PARZEN (1959)]. Let $m_1$ and $m_2$ belong to $\mathcal{H}(R)$. Then

$$E_{0,R} \left[ \frac{dP_{m_1,R}}{dP_{0,R}} \frac{dP_{m_2,R}}{dP_{0,R}} \right] = \exp(m_1, m_2)_{\mathcal{H}(R)}$$

where $E_{0,R}$ denotes expectation with respect to $P_{0,R}$.

THEOREM: UNEQUAL COVARIANCES [DUTTWEILER (1970)]. Let $R$, $R_1$ and $R_2$ be equivalent covariances. For $i = 1, 2$, $R_i \equiv R_0$ and $dP_{0,R_i}/dP_{0,R}$ is square integrable with respect to $P_{0,R}$ if and only if

$$\| R_i - R \|_{\mathcal{H}(R) \otimes \mathcal{H}(R)} \leq 1.$$  

When (2) holds for $i = 1, 2$, a formula for

$$E_{0,R} \left[ \frac{dP_{0,R_1}}{dP_{0,R}} \frac{dP_{0,R_2}}{dP_{0,R}} \right]$$

is given by Duttweiler in his thesis.

REFERENCES


2. A. Bensoussan, Sur l'identification et le filtrage de systèmes gouvernés par des équations aux dérivées partielles, Institut de Recherche d'Informatique et d'Automatique, Cahier no. 1, Rocquencourt, France, 1969.


