A View of Three Decades of Linear Filtering Theory

Invited Paper

THOMAS KAILATH, FELLOW, IEEE

Abstract—**Developments in the theory of linear least-squares estimation in the last thirty years or so are outlined. Particular attention is paid to early mathematical work in the field and to more modern developments showing some of the many connections between least-squares filtering and other fields.**

I. Introduction and Outline

THE SERIES of survey papers of which this is a part was begun largely to commemorate the twenty-fifth anniversary of the publication of Shannon's classic paper on information theory. However, 1974 is also twenty-five years after the publication in the open literature of Wiener's famous monograph, "Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications" [1], so that it is appropriate this year to commemorate this event as well. [As noted elsewhere in this issue, this month is also the tenth anniversary of Wiener's death (March 18, 1964).] Not only was this work the direct cause for the great activity of the last three decades in signal estimation, but it was perhaps the greatest factor in bringing the statistical point of view clearly into communication theory and also control theory. It may suffice to quote Shannon's own major acknowledgment: "Credit should also be given to Professor N. Wiener, whose elegant solution of the problems of filtering and prediction of stationary ensembles has considerably influenced the writer's thinking in this field."

The subject of estimation is a vast one, and most of our attention will be devoted to the particular problems of *linear least-squares estimation*, or *linear filtering* as it has generally come to be called in the engineering literature. Even though least-squares estimation is clearly only a small part of the possible forms of estimation theory, in the author's opinion it is perhaps the most interesting and most important part. Least-squares theory not only provides useful solutions to certain specific estimation problems, but it also has connections to and implications for a surprisingly large number of other problems, both statistical and deterministic. As some examples we mention signal detection [93], [291]; the calculation of mutual information in certain channels [303]; the solution of integral equations [288], [292], two-point boundary value problems in many fields; problems of scalar and matrix polynomial factorization with applications in network theory and stability theory [135], [177]; the solution of linear equations, especially as they arise in constructing state-variable realizations from impulse-response or transfer-function data, which in turn is related to the Berlekamp–Massey algorithm for decoding BCH codes [289], [191]; and the inversion of multivariable linear systems [357], [361], [362]. There are also more purely mathematical ramifications in Hilbert-space theory, operator theory, and more generally in functional analysis [245], [250], [261].

The section headings give a quick idea of the scope of the paper.

Needless to say, the choice of material and emphasis in this paper are mine; the field is a vast one and can be surveyed in various ways. My main aims are to provide some perspective on presently used methods, to bring out the significance and relevance of some relatively early, but often neglected, work in this field, and to illustrate some of the connections between least-squares theory and other fields.

In Section II we formulate the problem of determining the causal linear least-squares estimate of a signal process corrupted by additive white noise. Although this is only one of a large number of possible estimation problems, it is a key one in the sense that its solution underlies that of many others.

Sections III–VI describe some of the current approaches to solution of this key problem. Information theorists and communication engineers have been more familiar with problems in which covariance information is given about

Manuscript received June 15, 1973; revised October 3, 1973. This work was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, under Contract AF 44-620-69-C-0101, and in part by the Joint Services Electronics Program under Contract N-00014-67-A-0112-0044.

The author is with Stanford University, Stanford, Calif. 94305.

signal and noise, which are usually called Wiener filtering problems. Control engineers deal more often with problems where the signal and noise are described by state-space models, usually called Kalman filtering problems. Because of differing backgrounds, mutual knowledge of these two general approaches is often small, and one of our aims is to bring out the close and useful relation that must exist between these two approaches. Of course no proofs can be given here, but the main results are stated and their significance and role explained. Appropriate references are given for the proofs.

The discussion in Sections III–VI is fairly self-contained, but at various points allusions are made to earlier results in the mathematical literature, especially of the forties. This work is explored in Sections VII and VIII, partly for the record but really because it contains ideas that in my opinion have still not yet been adequately appreciated and exploited. For example, the work of Krein (1944) and of Levinson (1947) is only now beginning to be rediscovered and extended. Limitations of space again prevent any detailed discussion, but I have tried to provide some guidance for a reader interested in further study. Moreover, even a casual reader might find some fascinating nuggets exposed here, although I should stress that the lode is really much richer.

In Section IX, I have described in somewhat more detail one vein that I have personally found to be very illuminating and powerful: the role of canonical or innovations representations of random processes. Again I have given only references to many results and applications, but I could not resist being a little more specific about one aspect; namely, the connections to spectral factorization and to the so-called positive real functions of network and complex variable theory. The aim is to show at least one explicit connection between stochastic and deterministic problems. A fact I attempt to stress in Sections VIII and IX is the importance of deterministic system structure in the theory of random processes. This is at the moment an active field of research, but one that is being largely carried out in control theory. There is scope for many more communication theorists to enter this field. Conversely, in the last section I have briefly referred to some recent results in information theory that have useful implications for estimation. The time seems to be ripe for a fruitful symbiosis.

Section X is a brief look at the role and significance of series expansions. These should be relatively more familiar to readers of this journal, and I have therefore concentrated on some special, but often overlooked, aspects of such expansions. The section concludes in fact with an indication of how certain random process ideas can illuminate the currently active system-theoretic problem of abstract state-space determination. Again there is scope for fruitful interaction with control and system theorists. The paper closes with the thought that it is the range and scope of such possibilities that has kept estimation theory vital and active, without, it seems to me, the above-average doubts and misgivings that the more self-contained field of strict-sense Shannon theory has been exposed to recently.

The bibliography is of necessity rather vast, although it could easily have been even larger. On several occasions, the availability of convenient bibliographies has led me to omit various references. This is undoubtedly an injustice to the authors of many fine papers, but it seems to be unavoidable. I have attempted to do more justice to papers published in this journal, and in fact all such linear filtering papers in the period 1968–1972 have been included in the bibliography, even though no explicit reference may have been made to them in the text. This has also been done for certain other papers appearing in other journals that I feel contain some ideas or approaches that may appeal to our readers. The choice is of necessity rather subjective and any omissions should be regarded as a measure of my ignorance rather than a conscious slight.

The bibliography is organized under five subheadings, though the division of papers between the five categories is on occasion somewhat arbitrary, partly because the fields are of course not completely exclusive. In retrospect, some reassignments would really have been desirable, but I have not had the courage or the time to attempt them.

I must repeat that the inevitable limitations of time, space, and personal knowledge are undoubtedly reflected in this survey. The only palliative I can offer is that perusal of the various references will enable the reader to learn many additional facts and results that could not be covered in the paper and to make his own judgment of any controversial matters.

II. A Key Linear Estimation Problem

Some Early History

From the earliest times, people have been concerned with interpreting observations and making estimates and predictions. Neugebauer [370] has noted that the Babylonians used a rudimentary form of Fourier series for such purposes. As with so much else, the beginnings of a "theory" of estimation in which attempts are made to minimize various functions of the errors can apparently be attributed to Galileo Galilei in 1632 [174]. Then came a whole series of illustrious investigators, including the young Roger Cotes (of whom Newton said "had he lived, we might have known something"), Euler, Lagrange, Laplace, Bernoulli, and others.

As is well known, the method of least squares was apparently first used by Gauss in 1795 [197], though it was first published by Legendre in 1805 [198]. (It is less well known that Adrain in America, unaware of these developments, independently developed the method in 1808 [196]). Since then, there has been a vast literature on various aspects of the least-squares method. A comprehensive annotated bibliography of least-squares estimation for random variables has been given in a report by Harter [375]. (See also a brief survey by Sorenson [376].) Therefore we shall not go into this here, but shall proceed to least-squares estimation in stochastic processes, the first studies of which were made by Kolmogorov [207], [209], Krein [214], [215], and Wiener [1].

The works of Kolmogorov and Krein were independent of Wiener's, and while there was some overlap in the

results, their aims were rather different. Kolmogorov, inspired by some work of Wold [209], gave a comprehensive treatment of the prediction problem for discrete-time stationary processes. Krein noted the relationship of this work to some early work of Szegö [200], [201], on orthogonal polynomials, and extended the results to continuous time by clever use of a bilinear transformation. (We shall describe these results in more detail later (Section VII).)

However, no special attention was paid to explicit formulas for the optimum predictor itself. Such formulas are obviously necessary for applications, and in fact, certain anti-aircraft fire-control problems led Wiener to formulate independently the continuous-time linear prediction problem and derive an explicit formula for the optimum predictor. He also considered the "filtering" problem of estimating a process corrupted by a "noise" process. An interesting nontechnical account of this work and its background and development was given by Wiener in his autobiography [371, pp. 240–262].

Wiener constantly attempted to examine and stress the engineering significance of his ideas and results and his book [1] contains several explicit examples, which are still generally the only ones to be found in many textbooks on the subject. Wiener was also conscious of the problems of actually building circuits to implement the theoretical solutions. For example, he notes that "The detailed design of a filter involves certain choices of constants which must be justified economically. In general, it does not pay to eliminate a small error from a quantity when there is a large *irremovable error* in it." Another paragraph of his book is entitled "The Determination of Lag and Number of Meshes in a Filter." Partly because of such concerns, much of Wiener's work, despite its hard mathematics, has had a wide influence in engineering circles.

A Key Problem

As we begin to be more specific, we are immediately confronted with the fact that there is a large variety of estimation problems, even just linear ones. For example, as the reader knows, we can have prediction or filtering or smoothing problems, in state-variable form or transfer-function form, with additive white noise or colored noise, etc. However, in my opinion there is a *key* linear estimation problem in the sense that its solution can be shown to underlie that of many other problems (see [298], [181], for some examples): we have observations $y(\cdot)$ of a signal process $z(\cdot)$ in additive white noise $v(\cdot)$

$$y(s) = z(s) + v(s), \qquad t_0 \leq s \leq t_f \tag{1}$$

where

$$Ev(t)v'(s) = I_p\delta(t - s). \tag{2}^1$$

The usual assumption on $z(\cdot)$ is that it is uncorrelated with $v(\cdot)$

$$Ez(t)v'(s) \equiv 0. \tag{3}$$

[1] Some notational conventions: all random variables will be assumed to have zero mean. No special notation will be used to distinguish scalars and matrices; primes denote transposes; I_p is a $p \times p$ identity matrix. We could have assumed any strictly positive-definite matrix instead of I_p, but by a normalization we can return to the case described by (2).

However, this breaks down in feedback communication and feedback control problems where the signal $z(\cdot)$ may be influenced by past signal and noise. Therefore, a more general assumption is that

$$Ez(t)v'(s) = \begin{cases} \text{arbitrary}, & t \geq s \\ 0, & t < s. \end{cases} \tag{4}$$

It should be stressed that (4) can be introduced only with difficulty in many of the analyses found in the literature; for example, it cannot be directly handled by methods that rely on representing the signal and noise processes $z(\cdot)$ and $v(\cdot)$ by Karhunen–Loève expansions. (See Section X for further discussion of this point.)

To proceed with our description, let us define

$$K(t,s) = E[z(t)z'(s) + z(t)v'(s) + v(t)z'(s)]. \tag{5}$$

Note that $K(\cdot,\cdot)$ is generally not a covariance, unless $z(\cdot)$ and $v(\cdot)$ are uncorrelated. However $K(\cdot,\cdot)$ does determine the covariance function of the process $y(\cdot)$ as

$$R_y(t,s) = Ey(t)y'(s) = I_p\delta(t - s) + K(t,s). \tag{6}$$

We shall require that $R_y(t,s)$ be strictly positive definite on the square $[t_0,t_f] \times [t_0,t_f]$. Other assumptions will be that the signal process has finite expected energy and that $K(t,s)$ is continuous, though both of these assumptions can in fact be relaxed; the essential thing is really that $K(t,s)$ be "smoother" than $I_p\delta(t - s)$. (See also [327c].)

The problem is to determine a random variable $\hat{z}(t \mid t_f)$ of the form

$$\hat{z}(t \mid t_f) = \int_{t_0}^{t_f} H(t,\tau)y(\tau)\,d\tau, \qquad t_0 \leq t \leq t_f \tag{7}$$

so that

$$\operatorname{tr} E[z(t) - \hat{z}(t)][z(t) - \hat{z}(t)]' = \text{minimum}.$$

It is by now well known (see, e.g., [10]) that such an optimum linear least-squares estimate is characterized by the "orthogonality" property

$$E[z(t) - \hat{z}(t \mid t_f)]y'(s) = 0, \qquad t_0 \leq s \leq t_f \tag{8}$$

so that a simple calculation shows that the optimum filter $H(\cdot,\cdot)$ is determined by the solution of the integral equation

$$H(t,s) + \int_{t_0}^{t_f} H(t,\tau)K(\tau,s)\,d\tau = Ez(t)z'(s) + Ez(t)v'(s), \qquad t_0 \leq t, s \leq t_f. \tag{9}$$

Depending on whether $t_f < t$, $t_f = t$, or $t_f > t$, we have what are called *predicted*, *filtered*, or *smoothed* estimates, respectively. For convenience, we shall write $\hat{z}(t \mid t)$ as

$$\hat{z}(t) = \int_{t_0}^{t} h(t,s)y(s)\,ds.$$

Using (8) and (4), the relevant integral equation for *filtering* can be found as

$$h(t,s) + \int_{t_0}^{t} h(t,\tau)K(\tau,s)\,d\tau = K(t,s), \qquad t_0 \leq s \leq t \leq t_f. \tag{10}$$

This is a key equation in linear system theory.

It is important to note that (10) is substantially more difficult to solve than (9). Equation (9) is a Fredholm integral equation of the second kind and a lot is known about its properties. Several general solution methods are available, including reduction in different ways to a set of linear algebraic equations and the use of various gradient methods, see, e.g., [51], [37], [18], [27]. On the other hand, (10) is a much more difficult equation because of the constraint that $s \leq t$. This is seen most dramatically by taking

$$p = 1, \quad t_0 = -\infty, \quad t_f = +\infty, \quad K(t,s) = K(|t - s|)$$

in which case it is easy to see that in both (9) and (10) the solution $h(t,s)$ depends only on $|t - s|$. Then with an obvious change of variables (9) can be reduced to

$$H(t) + \int_{-\infty}^{\infty} H(\tau)K(t - \tau)\, d\tau = \tilde{K}(t), \qquad -\infty < t < \infty$$

which can be readily solved by Fourier transformation. However, the corresponding form of (10) is

$$h(t) + \int_{0}^{\infty} h(\tau)K(t - \tau)\, d\tau = K(t), \qquad 0 < t < \infty \quad (11)$$

which is quite a different thing. Simple Fourier transformation does not work, and we must in general use the celebrated Wiener–Hopf spectral factorization technique [2]–[4], which in fact has given (11) its name.

The Wiener–Hopf equation (11) first arose in astrophysics in 1894 and has been widely studied; see especially two comprehensive papers [234], [235], which deserve to be more widely read by engineers. On the other hand, much less is known about the nonstationary version (10), which we shall describe as being of Wiener–Hopf type.[2] Several references are collected in [257], [51], [298], and some of these will be brought up as we proceed with this survey. However, it is appropriate to begin with (11), which was where Wiener started.

III. Wiener Filters and Early Generalizations

The first explicit solutions for least-squares estimates of stochastic processes were given by Wiener in 1942 [1] under the assumptions of a scalar observation process ($p = 1$), a semi-infinite observation interval ($t_0 = -\infty$), and jointly stationary signal and noise processes. Wiener used a variational argument to determine the optimum estimate and was delighted to find that what was required was the solution of the Wiener–Hopf equation (11), a problem to which Hopf and he had ten years earlier contributed an elegant solution [2]. The method applies to quite general kernels, but, as Wiener himself noted, it took its simplest form for processes with rational spectral densities, for which $K(t,s)$ had the form

$$K(t,s) = K(|t - s|)$$

$$= \sum_{1}^{n} \alpha_i \exp(-\beta_i|t - s|)$$

$$= \begin{cases} \sum_{1}^{n} \alpha_i \exp(-\beta_i t) \exp(\beta_i s), & t \geq s \quad (12a)^3 \\ \sum_{1}^{n} \alpha_i \exp(-\beta_i s) \exp(\beta_i t), & t \leq s \quad (12b)^3 \end{cases}$$

where the $\{\beta_i\}$ are constants, possibly complex. For such kernels, the bilateral Laplace transform

$$S_y(s) = \int_{-\infty}^{\infty} [\delta(t) + K(t)] \exp(-st)\, dt \qquad (13)$$

is a ratio of polynomials in s^2, whose zeros display a quadrantal symmetry in the complex s-plane: every root of the form $\sigma + j\omega$ is accompanied by roots of the form $\sigma - j\omega$, $-\sigma \pm j\omega$. By virtue of this root distribution, we can make a *unique* factorization of $S_y(s)$ as

$$S_y(s) = S_y^{+}(s)S_y^{+}(-s) \qquad (14)$$

where the so-called *canonical factor* is

$$S_y^{+}(s) = \prod_{1}^{n} (s - z_i) \Big/ \prod_{1}^{n} (s - p_i) \qquad (15)$$

and $\{z_i\}$ and $\{p_i\}$ are the *left-half-plane* zeros and poles of $S_y(s)$, $\operatorname{Re} z_i < 0$, $\operatorname{Re} p_i < 0$. The Wiener–Hopf technique (see, e.g., [4], [20], [51]) shows that this factorization completely determines the Laplace transform of the optimum filter as

$$H_{\text{opt}}(s) = 1 - [S_y^{+}(s)]^{-1}. \qquad (16)$$

This expression, though implicit in Wiener's own examples, was first explicitly given by Yovits and Jackson [11] and by Krein [235].

Yovits and Jackson also gave a closed-form expression for the mean-square error in the special case of uncorrelated signal and noise

$$E[z(t) - \hat{z}(t)]^2 = \int_{-\infty}^{\infty} \ln [1 + S_z(i\omega)]\, d\omega. \qquad (17)$$

Since (17) does not require explicit knowledge of the optimum filter, it can be used to help decide if an optimum filter is worthwhile; several such applications in modulation theory have been discussed by Van Trees [40], [47], Stiffler [46], Lindsey [52], and the references therein. Recently Yao [48], [49], and Snyders [54], [54a], have extended this formula to cover certain problems with nonwhite noise, and signals or noise with nonrational spectra (see also Prouza [53]).

Wiener's theory has mainly been applied to the optimum choice of various components in modulation systems, as we have noted previously. We should explicitly mention

[2] One might argue that (10) is just a family of Fredholm equations, indexed by t. This is true, but the major problem lies in showing that the solutions $\{h(t,\cdot)\}$ can be satisfactorily fitted together, for example to make $h(\cdot,\cdot)$ square-integrable in both variables.

[3] This is not the most general form except when the poles $\{p_i\}$ in (15) are distinct; the generalization is not difficult but is notationally cumbersome and so has been avoided.

a notable early application to the design of loop filters in phase-locked loops [9]. Related applications were also made in optimal control (see [13], [24], and the references therein).

Some Generalizations

The Wiener–Hopf equation was soon extended to cover estimation of stationary processes given only over a finite observation interval ($t_0 > -\infty$), and more generally to cover the estimation of nonstationary processes. However, while it was not hard to discover that the general equation was of the form (10), there was no general method for solving such equations, and therefore a host of special results and techniques were developed. We especially mention an early paper by Zadeh and Ragazzini [6], which we shall encounter again in Section VIII. For processes with rational power-spectral densities, fairly explicit results were obtained by Yaglom [10], Hajek [243], Rozanov and Pisarenko [31], Whittle [32], Helstrom [33], and Slepian and Kadota [43].

Some solutions were also obtained for nonstationary processes [6], [7], [17], [21]. The most useful of these were by Shinbrot [14], [25], who however found it necessary to restrict attention to $K(\cdot,\cdot)$ of the form

$$K(t,s) = \begin{cases} \sum_1^n a_i(t)b_i(s), & t \geq s \quad (18a) \\ \sum_1^n a_i(s)b_i(t), & t \leq s. \quad (18b) \end{cases}$$

This is clearly a generalization of (12), but its true significance was not appreciated until later (see Section V). Despite such important contributions, however, too large a part of the literature dealt with minor variations and special cases, so much so that Elias felt compelled to editorialize in the IEEE TRANSACTIONS ON INFORMATION THEORY in 1958 that it was time to stop writing "two famous papers." One was "The Optimum Linear Mean-Square Filter for Separating Sinusoidally-Modulated Triangular Signals from Randomly-Sampled Stationary Gaussian Noise, with Applications to a Problem in Radar." (The other was "Information Theory, Photosynthesis, and Religion," a title suggested by D. Huffman.)

Furthermore, there were other reasons for being dissatisfied even with the most significant of the results of this period.

i) They were rather complicated, often requiring the solution of auxiliary differential and algebraic equations and the calculation of roots of polynomials.

ii) They were not easily updated with increases in the observation interval.

iii) They could not be conveniently adapted to the vector case ($p > 1$).

These last two difficulties came immediately to the fore in the late fifties in the problem of determining satellite orbits. Here there were generally vector observations of some combinations of position and velocity, and also there were large amounts of data sequentially accumulated with each pass over a tracking station. Swerling was one of the first to tackle this problem, and he presented some useful recursive algorithms [61], [80], [155], that were soon recognized and applied, especially by a group at the Bell Laboratories, who added various contributions of their own [84]. For different reasons, growing out of his successful application of state-space ideas in deterministic problems, Kalman developed [64], [68], [69], a somewhat more restricted algorithm than Swerling's, but it was one that seemed particularly matched to the dynamical state-estimation problems that were brought forward by the advent of the space age. Groups at the NASA Ames Laboratory [71], [72], and at the M.I.T. Draper Laboratories [70], [83], took up Kalman's ideas and developed them into programs that were successfully used in many space applications [139], [145], [138].

We shall examine the Kalman filter in some detail in Section IV. The reasons for our attention go beyond the specific algorithm and are more broadly connected with the importance of dynamical structure in data-processing algorithms. Unfortunately communication engineers and information theorists have lagged behind control engineers in appreciating this fact, though, as pointed out in Wong's recent survey [327], the gap is closing.

IV. KALMAN FILTERS

Kalman [64], [68], [69], changed the conventional formulation of the problem by giving, not the covariance of the signal process, but a "model" for it as the output of a dynamical linear system driven by white noise. Specifically, he assumed that the signal process $z(\cdot)$ could be described by

$$z(t) = H(t)x(t), \qquad t \geq t_0 \quad (19a)$$

$$\dot{x}(t) = F(t)x(t) + G(t)u(t), \qquad x(t_0) = x_0 \quad (19b)$$

where $x(\cdot)$ is an $n \times 1$ "state" vector and $u(\cdot)$ is an $m \times 1$ random input such that

$$Eu(t)u'(s) = Q(t)\delta(t - s) \quad (19c)$$

$$Ex_0x_0' = \Pi_0, \qquad Eu(t)x_0' \equiv 0, \qquad t \geq t_0. \quad (19d)$$

Also the matrices $F(\cdot)$, $G(\cdot)$, $H(\cdot)$, $Q(\cdot)$, and Π_0 are assumed known and continuous. In trajectory estimation (19b) could be the "linearized" equations of motion describing the evolution of the position and velocity vector $x(\cdot)$ subject to the wideband perturbations $u(\cdot)$ caused by random drag, gravitational uncertainties, etc., and the initial uncertainties x_0.

The assumption that x_0 and $u(\cdot)$ are uncorrelated not only is physically reasonable but also has the important consequence that the process $x(\cdot)$ is now a wide-sense Markov process [225]. Kalman also assumes that the "plant" noise $u(\cdot)$ and the "observation" noise $v(\cdot)$ in the observed process

$$y(t) = z(t) + v(t) = H(t)x(t) + v(t) \quad (20)$$

can be correlated, but he restricts the dependence to being of the form

$$Eu(t)v'(s) = C(t)\delta(t - s) \quad (21)$$

which is consistent with our more general earlier assumption (4) on the one-sided dependence between $z(\cdot)$ and $v(\cdot)$. The equations (19)–(21) describe the Kalman model for the estimation problem. The Kalman filter is not given by an explicit formula for the impulse response of the optimal filter, but as an algorithm suitable for direct evaluation by analog or digital computers

$$\hat{z}(t) = H(t)\hat{x}(t) \tag{22}$$

where

$$\dot{\hat{x}}(t) = F(t)\hat{x}(t) + K(t)\varepsilon(t), \qquad \hat{x}(t_0) = 0 \tag{23}$$

$$\varepsilon(t) = y(t) - \hat{z}(t) = y(t) - H(t)\hat{x}(t) \tag{24}$$

$$K(t) = P(t)H'(t) + G(t)C(t) \tag{25}$$

and the $n \times n$ matrix $P(\cdot)$ is the covariance matrix of the errors in the state estimates

$$P(t) = E\tilde{x}(t)\tilde{x}'(t), \qquad \tilde{x}(t) = x(t) - \hat{x}(t). \tag{26}$$

$P(\cdot)$ can be computed as the unique solution of the nonlinear differential equation

$$\dot{P}(t) = F(t)P(t) + P(t)F'(t) - K(t)K'(t) + G(t)Q(t)G'(t), \qquad P(t_0) = \Pi_0. \tag{27}$$

This equation is a matrix version of the familiar Riccati equation, first introduced by Francesco, Count Riccati, in 1724 [195] and since then often encountered in the calculus of variations. It seems surprising that a nonlinear equation should arise in a linear problem and be regarded as advantageous. However, the point is that it is a *differential equation* with known *initial* conditions, and such equations are comparatively easy to solve on a digital or analog computer because they involve only the iteration of relatively simple updating operations. This circumstance is indeed a very happy one, because Riccati equations can be introduced to solve general linear two-point boundary value problems, which arise often in various fields, see, e.g., [146], [247].

Discrete-Time Results

The discrete-time Kalman filter results were actually the first to be obtained, partly because the major system-theory activity in the mid-fifties was in the field of sampled-data systems, which arose when modern digital computers were put into control and communication links. Sampled-data filters for least-squares estimation were given by Franklin [8], Friedland [60], and others. While Friedland used infinite triangular matrices, Blum [59], [67], studied recursive filters for limiting the storage requirements of such algorithms. We have already noted the work of Swerling [61]. Kalman's contribution was to introduce state-space models. He assumed that

$$y_i = z_i + v_i, \qquad z_i = H_i x_i, \qquad i \geq 0 \tag{28a}$$

$$x_{i+1} = \Phi_i x_i + \Gamma_i u_i \tag{28b}$$

$$Eu_i x_0' \equiv 0 \equiv Ev_i x_0', \qquad Ex_0 x_0' = \Pi_0 \tag{28c}$$

$$Eu_i u_j' = Q_i \delta_{ij}, \qquad Ev_i v_j' = R_i \delta_{ij}, \qquad Eu_i v_j' = C_i \delta_{ij} \tag{28d}$$

where $\{\Phi,\Gamma,H,Q,R,C,\Pi_0\}$ are known matrices. The Kalman filter solution is

$$\hat{x}_{i+1|i} = \Phi_i \hat{x}_{i|i-1} + K_i(R_i^{\varepsilon})^{-1}\varepsilon_i, \qquad \hat{x}_{0|-1} = 0 \tag{29a}$$

$$\varepsilon_i = y_i - \hat{z}_{i|i-1}, \qquad \hat{z}_{i|i-1} = H_i \hat{x}_{i|i-1} \tag{29b}$$

$$R_i^{\varepsilon} = E\varepsilon_i \varepsilon_i' = H_i P_{i|i-1} H_i' + R_i, \qquad P_{i|i-1} = E\tilde{x}_{i|i-1}\tilde{x}_{i|-1}' \tag{29c}$$

$$K_i = \Phi P_{i|i-1} H_i' + \Gamma_i C_i \tag{29d}$$

where the $n \times n$ matrix

$$P_{i|i-1} \triangleq E[x_i - \hat{x}_{i|i-1}][x_i - \hat{x}_{i|i-1}]'$$

can be computed via the so-called Riccati difference equation

$$P_{i+1|i} = \Phi_i P_{i|i-1}\Phi_i' - K_i(R_i^{\varepsilon})^{-1}K_i' + \Gamma_i Q_i \Gamma_i', \qquad P_{0|-1} = \Pi_0. \tag{30}$$

The similarity of this set of equations to (22)–(27) is clear; in fact, the latter can be obtained from (28), (29), by a limiting procedure [68]. Note that because of the presence of the $P_{i|i-1}$-dependent term R_i^{ε}, the discrete-time formulas are somewhat more complicated than the continuous-time ones, or even than discretized versions of the continuous-time formulas. Moreover, in discrete time, there is no particular need to assume that the covariance of the additive-noise is nonsingular, and we have therefore written it as R_i rather than as I. Note that we could even take R_i to be zero without affecting the formulas (29), (30). This is not possible in continuous time, where problems with no nonsingular white noise component require more care (cf. [90], [122], [181] and the references therein). The study of state-estimation problems where there is no additive noise has recently uncovered some interesting differences between discrete- and continuous-time estimation and control problems (cf. [170], [179]).

By now the Kalman filter is widely known and widely used, notably in aerospace engineering; see, for example, the papers and references in the survey volumes [139] and [145]. Furthermore, (19)–(30) have turned out to have a fundamental role in understanding the structure and properties of dynamical systems, in many stochastic and deterministic problems. We may refer to work in quadratic optimization, stability theory, network theory, covariance and spectral factorization, stochastic control, sensitivity analyses, signal detection, etc. (The factorization problems will be briefly discussed in Section IX.) We forebear from giving specific references, but shall merely note some recent books in which such topics are covered [106], [128], [129], [132], [135], [143], [153], [157], [177]; it should be noted that Kalman himself launched the study of several of these questions.

However, despite all the good that has come out of this theory, there have been many excesses and oversights in its pursuit, partly reflected in an incredible volume of papers. Although some of this literature was necessary and worthwhile, a good fraction of it must be attributed to the general expansion of technological and especially space

activity that Sputnik and the Apollo project brought to America scene, in terms both of research and development contracts for industry and of the rapid expansion of graduate education in universities. Another factor was that this period coincided with the emergence of what is now called modern control theory, which was being built upon the rediscovery of the importance, emphasized in the mid-fifties by Bellman, of the notion of "state." The fact that the Kalman filter dealt largely with state-estimation made it comparatively easy to include it in books and courses in state-variable theory, without having to go very deeply into estimation theory or even into Wiener filtering.

Although several excellent examples of the clever and successful application of control theory ideas to estimation problems can be found, e.g., [74], [111], [128], [133], [148], [84a], [193a], the majority of contributions have suffered from having too narrow a base. I feel it is unfortunate that a whole generation of control engineers has grown up whose only knowledge of estimation theory is through Kalman filtering. The work of Kolmogorov, Krein, Wiener, Karhunen, Levinson, Lévy, Hida, and others (see Section VII) on many still important aspects of estimation has generally been neglected, not without loss. On the other hand, I should also state that in my opinion, the potential of the results and insights in the just-cited control-theoretic ideas has also not yet been fully exploited.

It may be useful to reinforce the previous comments by giving an illustration from control theory itself. As Kalman has often stressed [68] the major contribution of his work is not perhaps the actual filter algorithm, elegant and useful as it no doubt is, but the proof that under certain technical conditions called "controllability" and "observability," the optimum filter is "stable" or "robust" in the sense that the effects of initial errors and round-off and other computational errors will die out asymptotically. However, the known proofs of this result are somewhat difficult, and it is significant that only a small fraction of the vast literature on the Kalman filter deals with this problem. (Significant recent contributions have been made in [126], [158], [167], [151], [188].) The concepts of controllability and observability actually first arose as technical conditions in certain optimal control problems [119]. They also enter in a fundamental way [76], [119], in characterizing irreducible transfer functions and minimal state-space realizations of linear systems. Kalman isolated these notions and, for conceptual and other reasons, also defined them in terms of certain idealized but simple control problems; e.g., he observed that controllability is equivalent to being able to take an arbitrary initial state to the origin [62]. However, such definitions are only somewhat incidental, and their main justification lies in the theorems that can be proved with them. Nevertheless, many textbooks deal largely with examinations and elaborations of the definitions of controllability and observability, with hardly a mention of the associated theorems as being the reasons for this great attention. Information theorists may recognize similarities to the fate of the words information and entropy!

In my opinion, it was the peculiar atmosphere of the sixties, with its catchwords of "building research competence," "training more scientists," etc., that supported the uncritical growth of a literature in which quantity and formal novelty were often prized over significance and attention to scholarship. There was little concern for fitting new results into the body of old ones; it was important to have "new" results! Wiener had some interesting comments on the scene as early as 1956 [371, p. 271].

Despite this unfortunate historical context, one should not underestimate the significance of the Kalman filter, which, to repeat, is more than just a solution to a specific estimation problem. As a tribute to this work, I now attempt to add my view concerning the slight controversy that exists as to its origin.

Historical Notes on the Kalman Filter

Recursive solutions to least-squares problems are not of recent origin. Gauss was forced to invent them to handle the vast calculations he undertook in order to help astronomers locate the asteroid Ceres. His work dealt with the discrete-time model (28), where, however, the state x_i was constant (i.e., Φ and Γ were zero). Given hindsight one can generalize this work to handle dynamics and, for example, Rosenbrock has done this in his interesting note [92], (see also a note by Genin in [139]). Incidentally, Whittle in 1963 [32, p. 35] pointed out that the classical Wiener filter could be rewritten in a recursive form as a differential equation, and he also studied some nonstationary extensions [95].[4]

However, the general case was first studied by Kalman [64], who combined state-space descriptions and the notion of discrete-time innovations, as described for example in Doob [225, especially sects. XII.1 and XII.3], to give a complete and elegant solution. Kalman's solution also introduced a nonlinear recurrence equation (30) which was the discrete-time counterpart of a Riccati differential equation he had already encountered in studies on quadratic minimization problems in optimal control [63]. From this it was an easy step to obtain the continuous analog of the discrete-time equation for the least-squares estimate, especially since Kalman also recognized a "duality" between the filtering and control problems. An immediate bonus of his analysis of the steady-state behavior of the Riccati equation in optimal control was the important result that, under the previously mentioned technical conditions of "observability" and "controllability," the finite-time solution converges to a unique steady-state solution, independent of the initial condition and of errors introduced during the computation. [This stability question did not arise in the classical Wiener problem, which roughly speaking, corresponds to a Kalman filter problem with the $F(\cdot)$ matrix in the state-space signal model (19) constant and stable (i.e., having eigenvalues with negative real parts).

[4] Whittle [32], [95], used difference equation (autoregressive–moving average) models, and interestingly enough it is only recently that several advantages of such models have been fully appreciated (see the discussion at the end of Section IX).

Therefore, the signal variance goes to a steady-state value and it can be shown that so does the (always smaller) variance of the error. It is a striking fact, at least without further thought, that the error variance goes to a finite steady-state value under the structural conditions of controllability and observability even if F is unstable, so that the signal variance becomes unbounded.]

In view of these facts it seems fair to use the name Kalman filter for the continuous-time algorithm as well as for its discrete-time analog. The continuous-time filter is often also called the Kalman–Bucy filter, or sometimes the Bucy–Kalman filter (see Part I of [113] and also [136]). Bucy's coauthorship in Kalman and Bucy [69] grew out of some early work by Carlton and Follin [56] and Hanson [57], at the Applied Physics Laboratory of Johns Hopkins University, in which algorithms of the Kalman type were obtained for some special cases. Kalman's discovery of the general continuous-time formulas was apparently independent of this, being based as we have noted on analogies with optimal control [63]. Later Kalman obtained a direct derivation by applying a limiting argument to the discrete-time formulas [68]. Bucy's important contribution to the joint paper [69] was a derivation using the finite-time Wiener–Hopf equation (10). It should also be noted that Siegert in 1953–1955 [58] had already shown in a different context that finite-time Wiener–Hopf equations could be solved by reduction to a Riccati differential equation.

It is also not so widely known that, independently of all this, Stratonovich in the USSR had begun to study recursive solutions for nonlinear least-squares estimates of the states of a nonlinear dynamical system driven by white noise. In this connection, it is natural to consider the linearized problem and its solution, and in so doing Stratonovich in 1960 also obtained the Kalman filter equations ([65], [141, p. 675]). However, no stability analysis was undertaken.

We should also mention that with hindsight one can specialize certain recursive formulas obtained in 1958 by Swerling [61] for nondynamical systems to again obtain the Kalman filter. Swerling did not actually explicitly consider this special case, nor did he anywhere have a Riccati equation. However, as noted earlier, Swerling's papers [61], [155], contain several useful and interesting ideas, for linear and nonlinear filtering, many of which have been widely overlooked. [One such idea will be encountered in Section X.]

As a final comment on this topic, we may note that in a little known 1944 paper [213] (unfortunately not cited in his 1953 book, but added by Yaglom to the 1956 Russian translation) Doob made explicit and effective use of linear state-variable models to study processes with rational spectral density. This paper (see also [219]) contains several formulas and results that were rediscovered much later in the state-space literature.

V. Recursive Wiener Filters

Kalman replaced the conventional specification of the filtering problem in terms of signal and noise covariance functions by one in which state-space models were specified for the signal and noise, and it seemed to many that this difference in specification was the chief reason for the success of the Kalman filter. Therefore, it was thought that to obtain similar computationally efficient recursive solutions for problems with covariance information one should first deduce state-space models consistent with the given covariance specifications, to which the Kalman solution could then be applied. Unfortunately, the problem of determining such state-space models for nonstationary processes, which include stationary processes observed over finite time intervals, is quite difficult. Most known solutions require an amount of work roughly equivalent to that involved in solving the Riccati equation for a process with an already known signal model. Thus it appears in effect that the price paid for starting with covariance information rather than model information is essentially a doubling of work.

However, this is not true. With a proper formulation, the same amount of computation suffices to solve either problem. More specifically, suppose that we return to Shinbrot's covariance specification (18), which we shall rewrite more compactly in matrix notation as

$$K(t,s) = A(t)B(s)1(t - s) + B'(t)A'(s)1(s - t) \quad (31)$$

where $A(\cdot)$ and $B'(\cdot)$ are $p \times n$ matrices and $1(\cdot)$ is the Heaviside unit step function. The meaning of this assumption, which as stated earlier Shinbrot was forced to make for purely mathematical reasons, is that this is the form that $K(t,s) = E[z(t)z'(s) + z(t)v'(s) + v(t)z'(s)]$ must take for the processes $z(\cdot)$ and $v(\cdot)$ in a state-space model (19), (20).

Before proceeding, it will be convenient to rewrite $K(t,s)$ in the form

$$K(t,s) = M(t)\Phi(t,s)N(s)1(t - s) + N'(t)\Phi'(s,t)M'(s)1(s - t) \quad (32)$$

where $\Phi(\cdot,\cdot)$ is a so-called state transition matrix defined [106] as the unique solution of the linear differential equation

$$\frac{d\Phi(t,s)}{dt} = F(t)\Phi(t,s), \qquad \Phi(s,s) = I \quad (33)$$

and $F(\cdot)$ is an arbitrary matrix that can be chosen conveniently for the problem at hand. There is no loss of generality in doing this because $\Phi(\cdot,\cdot)$ is nonsingular and obeys $\Phi(t,s) = \Phi(t,t_0)\Phi(t_0,s)$ for an arbitrary t_0. When $F(\cdot)$ is constant

$$\Phi(t,s) = \exp F(t - s) \triangleq I + F(t - s) + F^2\frac{(t - s)^2}{2!} + \cdots. \quad (34)$$

For a *given* $F(\cdot)$, the correspondence between (29) and (30) is established by the relations (with t_0 arbitrary)

$$A(t) = M(t)\Phi(t,t_0), \qquad B(t) = \Phi(t_0,t)N(t)$$
$$M(t) = A(t)\Phi(t_0,t), \qquad N(t) = \Phi(t,t_0)B(t).$$

Now it is shown in [149] (see [292], [105], [144], for earlier efforts) that the estimate $\hat{z}(\cdot)$ can be calculated by the following recursive algorithm:

$$\hat{z}(t) = M(t)\phi(t) \tag{35a}$$

where

$$\dot{\phi}(t) = F(t)\phi(t) + K(t)[y(t) - M(t)\phi(t)], \qquad \phi(t_0) = 0 \tag{35b}$$

$$K(t) = N(t) - \Sigma(t)M'(t) \tag{36a}$$

$$\dot{\Sigma}(t) = F(t)\Sigma(t) + \Sigma(t)F'(t) + K(t)K'(t), \qquad \Sigma(t_0) = \mathbf{0}. \tag{36b}$$

The equation for $\Sigma(\cdot)$ is again an $n \times n$ nonlinear matrix differential equation of Riccati type. It is different from the Riccati equation (26) of the Kalman filter, though it is closely related [149]. As expected, (32)–(35) reduce to the Wiener filter formulas (11)–(16) in the special case $p = 1$, $t_0 = -\infty$, and $K(t,s)$ a function only of $|t - s|$. Thus we have now found the recursive generalization of the Wiener filter. To put it another way, we can now see how to make Shinbrot's integral equation solution recursive. A proof of these results is outlined in (128)–(133) of Section IX. (The discrete-time version of these equations can be found in [180].)

The important point is that the equations for $\Sigma(\cdot)$ and $\hat{z}(\cdot)$ can be directly written down from the covariance specifications without first having to determine a state-space model. Thus both specifications, in terms of covariances or in terms of state-space models, are seen to be equivalent, not just in that they give the same final answer (because, of course, they must), but in that their solutions involve comparable amounts of work. The choice between them lies purely in whether state-space models or covariance specifications are more readily at hand. This fact is still not widely appreciated and the literature contains many discussions of attempts to "identify" state-space models from covariance data so as to be able to use a Kalman filter.

Nevertheless, "modeling" is, as in all subjects, a thorny problem and we should say a few more words about it here. State-space models are often at hand in aerospace problems, where we may have enough information to write down the equations of motion, whether they be time invariant or time variant. However, the choice of the proper number of states to model a given problem adequately is not always an easy one. In many problems of industrial process control and communications, it is generally impossible to write down state equations (as is clear if we try to do so for a large power grid, or chemical plant or a telephone-line channel) and recourse has to be had to terminal measurements, for example of the covariance function or power spectrum of the channel output. Now covariance estimation is itself a vast subject, but even if we assume that good estimates are available, $K(t,s)$ will be available only as a numerical function of t and s and not in the factored form (31) or (32); getting the functions $\{A(\cdot),B(\cdot)\}$ or $\{M(\cdot),\Phi(\cdot,\cdot),N(\cdot)\}$ involves a further step of approximation. How can this be done? In the stationary case, where $A(\cdot)$ and $B(\cdot)$ are exponential functions, we have basically the classical network-theory problem of approximating a time function by a sum of exponential functions (or, in frequency-domain terms, approximating a function by a ratio of polynomials). A modern version of this problem is that of obtaining minimal state-space realizations from measurements of transfer functions. There are now several methods available for doing this, and in fact research into more efficient methods is still going on (see, e.g., [129], [360], and the references therein). However, this is only for the stationary case. Although certain analogous procedures can be devised for the nonstationary case [112], in general it is a difficult matter even in network theory to obtain time-variant system realizations.

For this reason, and also because attention is shifting away from aerospace problems, there is renewed interest in time-invariant models. Such models have generally been the only ones studied by statisticians and communication engineers, and recent research in control and system theory has taken a big swing in this direction.

VI. New Algorithms for Time-Invariant Systems

When the parameters in the Kalman state-space model are constant (time invariant), it has recently been discovered that one can obtain recursive solutions without going through a Riccati equation, and in several problems it is possible to obtain significant computational advantages thereby (cf. [166], [183], [184]). One reason for presenting this result is that a special case of it is closely related to algorithms invented in 1943–1947 by the astrophysicists Ambartsumian (USSR) [211] and Chandrasekhar (USA) [217], [220], to solve a class of Wiener–Hopf equations by reduction to nonlinear differential equations. This reduction was actually sought to obtain better numerical procedures, the same fundamental motivation underlying the work a decade later of Swerling, Stratonovich, and Kalman.

We shall start with the general state-space model (19), (20), where now F,G,H,Q,C are assumed to be time invariant. Then it has been shown [183] that the linear least-squares estimate of $z(\cdot)$ can be computed via the equations

$$\hat{z}(t) = H\hat{x}(t) \tag{37a}$$

$$\dot{\hat{x}}(t) = F\hat{x}(t) + K(t)[y(t) - H\hat{x}(t)], \qquad \hat{x}(t_0) = 0 \tag{37b}$$

which are (cf. (22), (23)) as in the Kalman filter, except that now $K(\cdot)$ need not be computed via a Riccati-type equation (27), but through the equations

$$\dot{K}(t) = L(t)SL'(t)H', \qquad K(t_0) = \Pi_0 H' + GC \tag{38a}$$

$$\dot{L}(t) = [F - K(t)H]L(t), \qquad L(t_0) = L_0 \tag{38b}$$

where S and the initial condition matrix L_0 are found as follows. Let

$$D \triangleq F\Pi_0 + \Pi_0 F' + GQG' - (\Pi_0 H' + GC)(\Pi_0 H' + GC)' \tag{39}$$

and suppose that

$$\text{rank of } D = \alpha, \qquad \alpha \leq n. \tag{40}$$

Since D is symmetric, we can write it via a standard numerical procedure (called the LDU decomposition, see e.g., [387]) as

$$D = L_0 S L_0' \tag{41}$$

where L_0 is an $n \times \alpha$ matrix and S is the $\alpha \times \alpha$ "signature" matrix of D

$$S = \text{diag}\,\{1,1,\cdots,1,-1,-1,\cdots,-1\}$$

with as many ones as D has positive eigenvalues. For reasons that will be given later, we shall say that the nonlinear differential equations (38) are of Chandrasekhar-type. This new algorithm determines $K(\cdot)$ directly via the solution of $n(p + \alpha)$ nonlinear differential equations for the $n \times p$ matrix $K(\cdot)$ and the $n \times \alpha$ matrix $L(\cdot)$. In the Kalman filter solution we have $n(n + 1)/2$ equations for the components of the matrix $P(\cdot)$, from which $K(\cdot)$ must then be found as $P(\cdot)H' + GC$.

In this generality, there may not be much to gain by choosing one algorithm or the other. However, there are important special cases where there is a decided advantage. For example, if we are very certain about the value of the initial state $x(t_0)$, we might assume that $\Pi_0 = 0$, which can be seen to lead to certain simplifications in D, in particular that $S = I$ and $\alpha \leq \min(n,m)$.

Another special case, of particular interest in communications and in system identification, is obtained by assuming that the state and signal processes $x(\cdot)$ and $z(\cdot)$ are in the statistical steady state, i.e., that they are stationary processes. To achieve this requires that we assume F to have eigenvalues with negative real parts and the initial state variance to be

$$\Pi_0 = \bar{\Pi}, \qquad 0 = F\bar{\Pi} + \bar{\Pi}F' + GQG'. \tag{42}$$

Now the matrix D reduces to

$$D = -(\bar{\Pi}H' + GC)(\bar{\Pi}H' + GC)' \text{ and rank } D \leq p. \tag{43}$$

Assuming for definiteness that the rank is p, the number of outputs we can take $S = -I$ and $L_0 = \bar{\Pi}H' + GC$, so that no special factorization is needed to specify the equations, which now comprise $2np$ nonlinear equations compared to $n(n + 1)/2$ for the Kalman filter. There can be a considerable computational saving when $n \gg p$.

If we assume not a state-space model for the *stationary* signal process $z(\cdot)$, but only knowledge of the covariance function of $z(\cdot)$, we shall be closer to the classical assumptions of the pre-Kalman theory, and will be able to bring out some interesting connections. Thus if (cf. (32))

$$R_y(t,s) = I_p\delta(t - s) + Me^{F(t-s)}N'1(t - s) + N'e^{F'(s-t)}M'1(s - t) \tag{44}$$

then it can be shown [183] that the algorithm is just (37) with (instead of (38))

$$\dot{K}(t) = -L(t)L'(t)H', \qquad K(t_0) = N \tag{45a}$$

$$\dot{L}(t) = [F - K(t)H]Y(t), \qquad L(t_0) = N \tag{45b}$$

which should be compared with the Riccati-type algorithm (35), (36) of the previous section.

We can now point out a close relationship to some famous equations obtained in astrophysics in connection with a Wiener–Hopf equation (10) with $K(t,s)$ of the form

$$K(t,s) = K(|t - s|) = \int_0^1 \exp(-|t - s|\alpha)w(\alpha)\,d\alpha \tag{46}$$

for a certain weighting function $w(\alpha)$. In 1947, Chandrasekhar [217] showed that the solution could be obtained in terms of two functions, now generally known as Chandrasekhar's X and Y functions, obeying the simultaneous nonlinear differential equations

$$\frac{\partial X(t,\alpha)}{\partial t} = -Y(t,\alpha)\int_0^1 Y(t,\beta)w(\beta)\,d\beta \tag{47}$$

$$\frac{\partial Y(t,\alpha)}{\partial t} = -\alpha Y(t,\alpha) - X(t,\alpha)\int_0^1 Y(t,\beta)w(\beta)\,d\beta \tag{48}$$

$$X(0,\alpha) = 1 = Y(0,\alpha), \qquad 0 \leq \alpha \leq 1. \tag{49}$$

These equations attracted considerable attention and have been studied, especially in transport theory and related fields, by Case [232], Noble [252], Wing [256], and Kalaba, Kagiwada, and Bellman (see [254] and the many references in the survey volume [146]). In 1972, Casti, Kalaba, and Murthy [161] used these results to show that the least-squares estimate $\hat{z}(\cdot)$ could be written

$$\hat{z}(t) = \int_0^1 L(t,\alpha)y(\alpha)\,d\alpha \tag{50}$$

where

$$\frac{\partial}{\partial t}L(t,\alpha) = -\alpha L(t,\alpha) + X(t,\alpha)[y(t) - \hat{z}(t)], \qquad L(0,\alpha) = 0, \qquad 0 \leq \alpha \leq 1. \tag{51}$$

With some small effort, the reader should be able to see that these equations are essentially the same as (37), (45a), (45b) if we make the assumptions

$$w(\alpha) = \sum_1^n \alpha_i\,\delta(\alpha - \alpha_i), \quad \alpha_i \geq 0, \quad -F = \text{diag}\,\{\alpha_1,\cdots,\alpha_n\}. \tag{52}$$

This is why the nonlinear differential equations (38) are said to be of Chandrasekhar type. We should note that the X and Y functions were already introduced by Ambartzumian in [211]. Chandrasekhar [217] first gave the differential forms, which Bellman, Kalaba, and their colleagues began to numerically exploit in the early sixties.

Discrete-Time Models

Analogous results can be obtained for discrete-time problems but the formulas are somewhat more complicated. This happened also with the Kalman formulas but the difference is even more pronounced here. Once again, we use the same estimator equation as in the Kalman filter

$$\hat{x}_{i+1|i} = \Phi\hat{x}_{i|i-1} + K_i[R_i^{\varepsilon}]^{-1}\varepsilon_i, \qquad \hat{x}_{0|-1} = 0 \tag{53a}$$

$$\varepsilon_i = y_i - H\hat{x}_{i|i-1} \tag{53b}$$

but K_i is found not via the Riccati difference equation (30), but via the equations [184]

$$K_{i+1} = K_i + \Phi L_i[R_i^r]^{-1}L_i'H' \tag{54a}$$

$$L_{i+1} = [\Phi - K_{i+1}[R_{i+1}^e]^{-1}H]L_i \tag{54b}$$

$$R_{i+1}^e = R_i^e + HL_i[R_i^r]^{-1}L_i'H' \tag{55a}$$

$$R_{i+1}^r = R_i^r - L_i'H'[R_i^e]^{-1}HL_i \tag{55b}$$

where

$$K_0 = \Phi_0 H' + \Gamma C, \qquad R_0^e = R + H\Pi_0 H' \tag{56}$$

and L_0, R_0^r are found by factoring the matrix

$$D \triangleq \Phi\Pi_0\Phi' + \Gamma Q\Gamma' - \Pi_0 - K_0[R_0]^{-1}K_0' \tag{57}$$

as

$$D = L_0 \begin{bmatrix} M_+ & 0 \\ 0 & M_- \end{bmatrix} L_0', \qquad M_+ > 0, \qquad M_- < 0. \tag{58}$$

The matrix L_0 has dimension $n \times \alpha$, where α = rank of D. Then L_0 is the initial value for (54b), while

$$R_0^r = \begin{bmatrix} M_+^{-1} & 0 \\ 0 & M_-^{-1} \end{bmatrix}. \tag{59}$$

The form of (59) suggests that we define

$$M_i = [R_i^r]^{-1}$$

in which case (55b) gives the equation

$$M_{i+1} = M_i + M_iL_i'H'[R_i^e]^{-1}HL_iM_i. \tag{60}$$

We could obtain other variations by updating $[R_i^e]^{-1}$ directly. The best choices seem to be those we have given; a count of the number of operations, which is more significant in discrete time than the number of equations, shows that all forms of the new algorithm involve less computation than the Kalman filter. Recently, other variants have been found in which specific equations as in (54) or (60) are replaced by the specification of successive orthogonalizing transformations (e.g., of the Householder type) to be applied to certain data arrays [193c], [193d]. These forms are intimately related to square-root estimation algorithms (see [98a], [128a], [148], and the references therein), and to the ideas of canonical spectral factorization (shades of Wiener, again). This last topic is further discussed at the end of Section IX.

As in continuous time, in various special cases the algorithm can be simplified further, e.g., when $\Pi_0 = 0$ or when $\Pi_0 = \bar{\Pi} = \Phi\bar{\Pi}\Phi' + \Gamma Q\Gamma'$. In the latter case, the processes z and y are stationary, and it can be shown that the relevant equations are closely related to the work of Levinson in 1947 [218]. Curiously, the algorithm for the continuous-time stationary process case is also related to work done in 1947, namely that of Chandrasekhar [217]. While Chandrasekhar's paper was in astrophysics and therefore somewhat inaccessible, it is unfortunate that Levinson's paper, which was reprinted as an appendix in Wiener's monograph [1], has been somewhat overlooked in the literature of communication and control theory, though it has been widely used by geophysicists [79], [96], [110], and very recently in speech analysis [164], [173]. Since they solve similar problems, there must clearly be a close connection between Chandrasekhar's and Levinson's results. Actually, both can be derived by using certain "invariance" principles, which consist basically of "invariantly imbedding" the given problem in a family of similar problems. This will be explained in some detail in Section VII.

We should also note that the techniques used to obtain (38)–(41) and (54)–(58) can also be applied [183], [184], to other problems where Riccati equations arise, even to general two-point boundary value problems whose solution is well known to be obtainable via a nonsymmetric Riccati equation. Furthermore, by using the ideas discussed at the end of Section IX, the results can also be extended to certain classes of time-variant and even nonlinear models. Applications to infinite-dimensional (distributed parameter) problems seem to hold special computational promise since operator Riccati equations on Hilbert $\times$ Hilbert spaces are replaced by equations on Hilbert $\times$ R^n spaces.

As stressed in [184], [193d], a significant aspect of all the results of this section is that a reevaluation is timely of the almost total concentration on the Riccati equation in the sixties. The rest of this paper hopes to describe some of the concepts that will underlie such a reexamination. These concepts have actually been available for quite a while, but, as stated before, they seem to have been generally neglected, perhaps as historical curiosities. However, many of the results of Sections V and VI would probably not have been developed without awareness of the important role of the Wiener–Hopf equation and of spectral factorization in this field. Our survey of these various ideas can only be partial, but it is deliberately also somewhat tutorial so as to aid the interested reader in making a closer study of the many references that will be noted later.

VII. Some Early Mathematical Work on Linear Least-Squares Estimation

The adjectives in the title might seem strange to someone who has gone through the previous sections; it might also seem a bit presumptuous considering how often the work of Wiener has been mentioned so far. Nevertheless, as stated by Masani [255] in the special Wiener issue of the Bulletin of the American Mathematical Society, the portion of Wiener's work [1] that we have described does not have "the theoretical strength and completeness of that of Kolmogorov." Wiener became aware of this himself when he tackled the multivariate and nonlinear least-squares problems. Masani writes that "Wiener adopted the [Kolmogorov] Hilbertian approach in his later papers under the stimulus of his younger collaborators."

Before embarking upon our examination of the more mathematical work of Kolmogorov and his successors, we should perhaps reassure the reader that the mathematical level of our presentation is not going to take a big jump. "Deeper" mathematics, or even "more abstract" mathematics, does not necessarily entail more formidable mathematical "language." It is quite possible to present the basic ideas of the deeper mathematics in a physical way, and in fact it is no longer a novelty that often the deeper mathematics is closer to physical constructs. The

Schwartz theory of generalized functions is a good example, where the abstract topological notions of generalized equality and convergence are more closely tied to the mechanisms of physical measurements than are the classical pointwise or L^p definitions.

The Work of Wold (1938)

In 1938, just a few years after Kolmogorov had put the theory of probability and stochastic processes on a sound general footing [203], Wold presented a Ph.D. dissertation on discrete-time stationary processes. This dissertation, now available in book form [206], contains several interesting results, of which we mention only a very few, related to our theme. For example, Wold already used the idea proposed by Frechet in 1937 [205] of regarding random variables as elements of a metric space with the distance between two elements being the variance of their difference. This geometric interpretation made it natural to interpret least-squares estimation as projection onto a subspace. It took many years for this natural idea to penetrate the engineering literature, where even in the sixties, strenuous efforts were made in many papers to avoid using the so-called "orthogonality" condition for least-squares estimates. We may mention in passing that Wold was influenced by the work of Frisch [202], where as Wold states, "matrix calculus was for the first time systematically employed in statistics." In 1969 Frisch received the first Nobel Prize in Economics.

One of Wold's major observations was that it simplified calculations to replace a sequence of correlated random variables by an "equivalent" sequence of uncorrelated variables. He also noted that certain processes could be "singular" in that their future values could be predicted exactly from knowledge of their past values. Such processes are nowadays, following Doob [213], called *deterministic* processes. These various ideas were combined into the following fundamental result [206, p. 89].

> Let $y(t)$ be a finite-variance stationary discrete-time process. Then there exist three *jointly stationary* processes $\{\chi(t),\varepsilon(t),\psi(t)\}$
>
> i) $y(t) = \chi(\cdot) + \psi(\cdot)$;
> ii) $\psi(\cdot)$ and $\chi(\cdot)$ are uncorrelated;
> iii) $\psi(\cdot)$ is deterministic and unique;
> iv) $\varepsilon(\cdot)$ has uncorrelated components, $E\varepsilon(i)\varepsilon(j) = \delta_{ij}$;
> v) $\chi(t) = b_0\varepsilon(t) + b_1\varepsilon(t-1) + b_2\varepsilon(t-2) + \cdots$, where $\sum b^2 < \infty$.

This decomposition is now called the *Wold decomposition* and it has been widely used and generalized, especially in functional analysis [245], [250], [261].

The Work of Kolmogorov (1939–1941)

While Wold went on in his thesis to apply his ideas to economic time series, it was left to Kolmogorov to pick up and complete Wold's results on prediction, which he did in a brilliant and comprehensive fashion in the papers [207], [209], [210]. Later Wiener made similar, but less complete efforts, and Masani [255] comments that "so thorough had been Kolmogorov's treatment of univariate prediction in the discrete case that there was little left to do."

Kolmogorov first noted that, though Wold's decomposition was stated as an existence theorem, it becomes a prediction formula as soon as the deterministic process $\{\psi(\cdot)\}$ and the coefficients $\{b_i\}$ of the moving-average process (so named by Kolmogorov $\{\varepsilon(\cdot)\}$) are fixed. Thus suppose we know that $\psi(\cdot)$ is identically zero. Wold was aware, but did not explicitly state in his theorem, that $\varepsilon(t)$ could be uniquely determined by $\{y(t), y(t-1),\cdots\}$. In fact, Wold essentially constructed the $\varepsilon(\cdot)$ sequence by successive Gram–Schmidt orthogonalization, for which this property is obvious. (We say "essentially" because Wold was dealing with an infinite sequence $\{y(t), y(t-1),\cdots\}$ and a possibly nonzero deterministic $\psi(\cdot)$, so that a careful double limiting procedure had to be used.) This property was explicitly introduced and exploited by Kolmogorov as follows. By Wold's theorem

$$y(t) = b_0\varepsilon(t) + b_1\varepsilon(t-1) + b_2\varepsilon(t-2) + \cdots \quad (60a)$$

where the $\{\varepsilon(t), \varepsilon(t-1),\cdots\}$ are uncorrelated random variables that can be computed from $\{y(t), y(t-1),\cdots\}$ by linear operations. Also

$$y(t+1) = b_0\varepsilon(t+1) + [b_1\varepsilon(t) + b_2\varepsilon(t-1) + \cdots].$$

However, as just noted, the terms in the square bracket are completely determined by knowledge of past $y(\cdot)$; moreover, these terms are not correlated with $\varepsilon(t+1)$. Therefore, we have

$$\hat{y}(t+1 \mid t) \triangleq \text{linear least-squares estimate of } y(t+1) \text{ given } \{y(t), y(t-1),\cdots\}$$
$$= b_1\varepsilon(t) + b_2\varepsilon(t-1) + \cdots.$$

This solves the prediction problem. Here

$$b_0\varepsilon(t+1) = y(t+1) - \hat{y}(t+1 \mid t)$$

so that one may call $\varepsilon(t+1)$ the "new information" or the "innovation" in the process $y(\cdot)$ at time $t+1$, and the process $\varepsilon(\cdot)$ may be called the *innovations process* of $y(\cdot)$, a name that was apparently first used by Wiener and Masani in the mid-fifties (personal communication in 1968 from P. Masani). (See also a 1960 paper by Cramer [268].) We shall see that such processes play a fundamental role in our understanding of the process $y(\cdot)$ in both discrete and continuous time (Sections VIII and IX).

So far we have chiefly an easy application of Wold's theorem. Kolmogorov went on to deepen Wold's theorem by relating it to properties of the so-called integrated power spectrum of the process $y(\cdot)$. With the covariance

$$R(i-j) \triangleq Ey(i)y(j)$$

Wold had shown that there exists a nondecreasing function $F(\lambda)$ called the *integrated power spectrum* of $y(\cdot)$ such that

$$R(k) = \int_{-1/2}^{1/2} \exp(i2\pi k\lambda)\, dF(\lambda).$$

In general $F(\cdot)$ will consist of an absolutely continuous part, a jump part, and a singular part (continuous and nondecreasing but with zero derivative almost everywhere). Kolmogorov showed that the deterministic part of a process $y(\cdot)$ is identically zero if and only if $F(\lambda)$ is absolutely continuous and its derivative $\dot{F}(\lambda)$ satisfies

$$\int_{-1/2}^{1/2} \ln \dot{F}(\lambda)\, d\lambda > -\infty. \tag{61}$$

He also gave explicit formulas for the coefficients $\{b_i\}$ in terms of the Fourier series coefficients of $\ln \dot{F}(\lambda)$[5] and finally showed that the one-step prediction error has a simple form

$$\lim_{t\to\infty} E[y(t) - \hat{y}(t \mid t-1)]^2 = \lim_{t\to\infty} E\varepsilon^2(t) = \exp \tfrac{1}{2} \int_{-1/2}^{1/2} \ln \dot{F}(\lambda)\, d\lambda. \tag{62}$$

This last formula, which is valid for all processes (even those with a nonzero deterministic part), is closely related to the Yovits–Jackson formula (17) mentioned in Section III (cf. [48], [54]).

These remarkable results were obtained by connecting the study of stationary random processes with that of certain deterministic functions in the frequency domain. Kolmogorov did this via the relationship

$$Ey(k)y(l) = R(k-l) = \int_{-1/2}^{1/2} \exp i2\pi\lambda(k-l)\, dF(\lambda). \tag{63}$$

It will be useful later to set

$$z = \exp i2\pi\lambda$$

so that (63) can also be written

$$Ey(k)y(l) = R(k-l) = \oint z^k z^{-l} \frac{dF(z)}{z} \tag{64}$$

where the special symbol denotes integration around the unit circle. The left side of (63) can be regarded as an inner product between the random variables $y(k)$ and $y(l)$ in a Hilbert space of random variables formed from finite linear combinations of the $\{y(k)\}$ and their limits in the covariance norm. Similarly, because $F(\lambda)$ is nondecreasing, the right side can be regarded as an inner product between $\exp i2\pi k\lambda$ and $\exp i2\pi l\lambda$ in the space $L_2(dF)$ of functions square-integrable with respect to the measure $dF(\lambda)$. Similarly the right side of (64) defines an inner product between the powers z^k in the space of polynomials on the unit circle. Therefore, we have an obvious norm-preserving mapping (or isometry) between these spaces in which

$$y(k) \leftrightarrow \exp i2\pi k\lambda \leftrightarrow z^k \tag{65}$$

$$\sum_1^n a_k y(k) \leftrightarrow \sum_1^n a_k \exp i2\pi k\lambda \leftrightarrow \sum_1^n a_k z^k. \tag{66}$$

[5] Incidentally, this defines the so-called "cepstrum" of $y(\cdot)$, which has had a vogue in signal processing recently [364]–[368].

These λ-functions can be extended from the interval $(-\frac{1}{2},\frac{1}{2})$, or equivalently from the unit circle, into the complex plane, an idea of Hardy's [199] that had been extensively studied by Paley and Wiener [204]. Nevertheless, it was Kolmogorov who exploited these ideas in general prediction.

A Formula of Szegö's (1915)

Kolmogorov's explicit formula (62) for the mean-square error had already been discovered in a different, but isomorphic problem by G. Szegö in 1915 [200], [201]. The isomorphism was precisely the one introduced by Kolmogorov, according to which the problem of choosing $\{a_j\}$ to minimize

$$\sigma_n^2 \triangleq E\left[y(n) - \sum_0^{n-1} a_j y(j)\right]^2 \tag{67}$$

is the same as that of choosing them to minimize

$$\sigma_n^2 = \int_{-1/2}^{1/2} \left|\exp(i2\pi\lambda n) - \sum_0^{n-1} a_j \exp(i2\pi\lambda j)\right|^2 dF(\lambda) \tag{68}$$

or

$$\sigma_n^2 = \oint \left|z^n - \sum_0^{n-1} a_j z^j\right|^2 \frac{dF(z)}{z}. \tag{69}$$

Thus the problem of minimizing σ_n^2 by suitable choice of the $\{a_i\}$ is just a problem of *polynomial approximation* on the unit circle. This problem was solved for absolutely continuous $F(z)$ by Szegö, and rederived by Kolmogorov for general $F(z)$. The connection to Szegö's work was noted by Krein [214] and later by Grenander [223].

The Work of Krein (1944–1945)

In response to questions raised by Kolmogorov, Krein in 1944–1945 [214], [215], showed how Kolmogorov's results could be extended to continuous time by use of a simple bilinear transformation. To each *discrete-time* stationary process with integrated spectrum $F(\lambda)$ we can associate a stationary *continuous-time* process with integrated spectrum $S(f)$, where

$$F(\lambda) = S(f), \qquad \lambda = \frac{1}{\pi}\tan^{-1} f \tag{70}$$

$$\exp i2\pi\lambda = \frac{1+if}{1-if}, \qquad \tan \pi\lambda = \frac{\exp(2\pi i\lambda) - 1}{\exp(2\pi i\lambda) + 1} = if. \tag{71}$$

This transformation has the useful property of preserving causality, and therefore it is often used in digital signal processing (see, e.g., [367]).

Use of the bilinear transformation shows easily [225, ch. XII] that the necessary and sufficient condition for no deterministic part is (compare (64))

$$\int_{-\infty}^{\infty} \frac{\ln \dot{F}(f)}{1+f^2}\, df > -\infty. \tag{72}$$

The bilinear transformation can also be extended to multivariable systems [233] and to systems in state-space form

[163],[6] where it has been exploited to give a new technique for solving the steady-state Riccati equation. The bilinear transformation was also used in some generality by Masani and Robertson [246], [35], [45], who paid particular attention to the question of how the discrete-time innovations process goes over to continuous time. We may note that the discrete-time one-step prediction error formula has no continuous-time analog, but there is a continuous-time version of the Wold decomposition and of the innovations process $\varepsilon(\cdot)$. These are important results, which we shall discuss further in the next section.

Krein has made several other important contributions to filtering theory. In 1954, he discovered that the spectral analysis of a weighted string, in which he had been interested since 1940, enabled him to obtain [227] some deep results on estimation given a finite data segment. Krein's analysis also led him to several other results on the solution of integral equations, the so-called "inverse-scattering" problem (see, e.g., [224], [226]). Recently, Dym and McKean [259], [260] have pursued Krein's ideas even further.

The Work of Levinson (1947)

In the USA, work on least-squares estimation proceeded along different lines. Wiener's basic ideas, rather than his ingenious solution of the problem, influenced work on anti-aircraft devices, where the need to find computationally simple solutions led to some alternative approaches. Thus Phillips [212] began with the assumption that the optimum filter has a rational transfer function and used the mean-square-error criterion to solve for the coefficients. Others, apparently including Blackman, Bode, and Shannon [3], tried to incorporate the dynamical constraints on the targets into the prediction schemes.

Levinson [218] formulated the problem in discrete time. In his words, "A few months after Wiener's work appeared, the author, in order to facilitate computation procedure, worked out an approximate, and one might say, mathematically trivial procedure." Levinson's deprecatory comments notwithstanding, this work has had an important impact on the field, both directly and indirectly. His equations were rediscovered in 1960 by Durbin [239] in a scheme for recursive fitting of autoregressive models to scalar time-series data. Whittle [81], [32] extended these recursions to multivariate time series, and his work has been widely used by statisticians. Levinson's work was directly used and extended to multivariate series by workers in geophysics, especially Robinson, with various contributions by groups in the Geology Department at M.I.T. and in the oil industry (cf. [110] and the references therein). These algorithms are now being used in speech analysis (see, e.g., [164], [173]) and in spectral estimation [107], [156], [182]. There are also close relations to the theory of orthogonal polynomials [208], [236], [241] and to the algorithms of Section VI. These and other connections will be briefly discussed after we present the basic problem and its solution.

Given a segment of a stationary time-series $\{y(0), y(1),\cdots, y(N-1)\}$, where the $\{y(i)\}$ are p-vectors, we wish to find the optimum one-step prediction

$$\hat{y}_{N|N-1} \triangleq -\sum_{i=0}^{N-1} A_{N,N-i}y(i). \tag{73}$$

Let

$$R_{i-j} \triangleq Ey(i)y'(j)$$

then by using the orthogonality property of least-squares estimates, we will have the equations

$$R_{N-j} = -\sum_{i=0}^{N-1} A_{N,N-i}R_{i-j}, \qquad j = N-1,\cdots,0. \tag{74}$$

The mean-square error is given by

$$R_N^{\varepsilon} \triangleq = E[y_N - \hat{y}_{N|N-1}]y_N' = R_0 + \sum_{0}^{N-1} A_{N,N-i}R_{i-N}. \tag{75}$$

Since R_N^{ε} is a nonincreasing function of N, its value can be used to decide whether it is necessary to collect more data (i.e., increase N) in order to achieve a desired mean-square error. As stressed by Levinson, this makes it important to find a way of successively calculating R_N^{ε}, $N = 0,1,\cdots$. The first step is rearrange the filter equations (74) and the error equation (75) in a single block-Toeplitz matrix equation

$$[I,A_{N,1},\cdots,A_{N,N}]\underbrace{\begin{bmatrix} R_0 & R_1 & \cdots & R_N \\ R_{-1} & R_0 & \ddots & \\ \vdots & \ddots & \ddots & R_1 \\ R_{-N} & \cdots & R_{-1} & R_0 \end{bmatrix}}_{\mathscr{R}_N} = [R_N^{\varepsilon},0,\cdots,0] \tag{76}$$

where $\mathscr{R}_N$ is a block-Toeplitz matrix. The unknowns are the $\{A_{N,i}\}$ and R_N^{ε}. The aim is to determine R_{N+1}^{ε} and $\{A_{N+1,i}\}$ in a way that takes maximum advantage of the previous computations made to find R_N^{ε} and $\{A_{N,i}\}$. It takes almost as long to describe the result as it does to give a derivation, following Robinson [110, ch. 6]. We shall therefore do so here, partly also with the hope that readers may recognize analogies with similar procedures in other problems.

The method is first to try an "obvious" solution, pushing our luck the most by assuming that adding a zero to the previous solution may work. It will if in the resulting equation

$$[I,A_{N,1},\cdots,A_{N,N},0]\mathscr{R}_{N+1} = [R_N^{\varepsilon},0,\cdots,0,\alpha_N] \tag{77}$$

the term $\alpha_N = R_{N+1} + \sum_1^{N-1} A_{N,i}R_{N-i}$ is zero. If this does not happen, we have to find a (simple) way of forcing α_N to zero. For this we introduce the "auxiliary" (adjoint or reversed) equations

$$[0,B_{N,N},\cdots,B_{N,1},I]\mathscr{R}_{N+1} = [\beta_N,0,\cdots,0,R_N^r] \tag{78}$$

where at stage N we assume that we know $\{A_{N,i},R_N^{\varepsilon},\alpha_N\}$ and $\{B_{N,i},R_N^r,\beta_N\}$. [For $N = 1$, we take $B_{1,1} = I$, $R_0^r = R_0 = R_0^{\varepsilon}$.] Next we form a weighted combination of (77),

[6] See also Popov [87], [87a], papers that contain several important ideas on spectral factorization and innovations representations (see especially [87, sects. 5, 7, and appendix E, F).

(78)

$$[I,A_{N,1} + K_N^{\alpha}B_{N,N},\cdots,K_N^{\alpha}]\mathcal{R}_{N+1} = [R_N^{\varepsilon} + K_N^{\alpha}\beta_N,0,\cdots,0,\alpha_N + K_N^{\alpha}R_N^{r}]$$

from which it is easy to see that choosing

$$K_N^{\alpha} = -\alpha_N[R_N^{r}]^{-1} \tag{79}$$

gives us a solution of the extended equation, i.e.,

$$[I,A_{N+1,1},\cdots,A_{N+1,N+1}] = [I,A_{N,1},\cdots,A_{N,N},0] + K_N^{\alpha}[0,B_{N,N},\cdots,B_{N,1},I]. \tag{80}$$

Similarly with the help of

$$K_N^{\beta} = -\beta_N[R_N^{\varepsilon}]^{-1}$$

we can update the $\{B_{N,i}\}$ as

$$[B_{N+1,N+1},\cdots,B_{N+1,1},I] = [0,B_{N,N},\cdots,B_{N,1},I] + K_N^{\beta}[I,A_{N,1},\cdots,A_{N,N},0]. \tag{81}$$

These relations were independently discovered by Whittle [81] and Wiggins and Robinson [96]. However, the latter had the benefit of an observation of Burg that

$$\alpha_N = \beta_N'. \tag{82}$$

The recursions for R_N^{ε} and R_N^{r} can be seen to be

$$R_{N+1}^{\varepsilon} = R_N^{\varepsilon} - \alpha_N[R_N^{r}]^{-1}\beta_N \tag{83}$$

$$R_{N+1}^{r} = R_N^{r} - \beta_N[R_N^{\varepsilon}]^{-1}\alpha_N. \tag{84}$$

Some Connections and Extensions

We should point out that the method of using a partial solution and a subsequent correction using a "reversed-time" auxiliary solution is the same in spirit as the method of invariance introduced into astrophysics by Ambartsumian [211] and Chandrasekhar [217]. This is a powerful idea, whose main thrust can be recognized in many problems—e.g., in the method of adjoints (or influence functions) in optimization problems [128] or in the Berlekamp–Massey decoding algorithm for BCH codes [289], [354], to mention only a few problems that may be of interest to readers of this article (see also [191]).

Here we note the obvious similarity between (54), (55) and (83), (84). In [185] we have given a stochastic interpretation of the previous algorithm, in which the Burg identity (82) arises very naturally. These ideas are then extended to a general study of decompositions of a time-series (into *conditional innovations* and *residuals*) that yield a certain statistical shift-invariance property. This property, analogous in many ways to the natural shift-invariance of stationary processes, enables "imbedding" derivations of the discrete-time algorithms (53)–(60) of Section VI and of certain so-called "Fast-Cholesky" algorithms for triangularizing nonnegative-definite block-Toeplitz and certain related matrices [191], [193].

Orthogonal Polynomials

It is natural to describe the relations (80)–(83) in polynomial language. Let

$$A_N(z) = A_{N,N} + \cdots + A_{N,1}z^{N-1} + z^N \tag{85}$$

where z is an indeterminate and similarly define $B_N(z)$, $A_{N+1}(z)$, $B_{N+1}(z)$. Then the recursions (81)–(83) can be written compactly as

$$\begin{bmatrix} A_{N+1}(z) \\ B_{N+1}(z) \end{bmatrix} = \begin{bmatrix} I & zK_N^{\alpha} \\ zK_N^{\beta} & I \end{bmatrix} \begin{bmatrix} A_N(z) \\ B_N(z) \end{bmatrix} \tag{86}$$

which turn out to be exactly the recurrence formulas for orthogonal polynomials on the unit circle [236], [241]. These are polynomials $\{A_N(z), |z| = 1\}$ such that

$$\langle A_i(z),A_j(z)\rangle \triangleq \oint A_i(z)\frac{dF(z)}{z}A_j'(z) \tag{87}$$

$$= \delta_{ij}R_i^{\varepsilon} \tag{88}$$

where $F(\cdot)$ is the matrix integrated power spectrum (cf. (63)) defined by

$$\oint z^i z^{-j}\frac{dF(z)}{z} = \text{a covariance function, } R_{i-j}. \tag{89}$$

A striking thing about these polynomials is that, unlike the classical orthogonal polynomials on the line, to get a nice recursion one has to introduce "auxiliary" polynomials $\{B_i(z)\}$ and define a simultaneous recursion for the $\{A_i(z)\}$ and the $\{B_i(z)\}$.

This fact was first noticed by Szegö [200] who considered the scalar case in which the previous relations simplify because the Toeplitz matrix $\mathcal{R}_N$ will be symmetric rather than just block-symmetric. Now it can be seen that

$$B_N(z) = z^N A_N(z^{-1}) = z^N + A_{N,N}z^{N-1} + \cdots + A_{N,1} = \text{the "reverse" of the polynomial } A_N(z). \tag{90}$$

In 1961, Baxter [240] extended Szegö's formulas to scalar nonsymmetric Toeplitz matrices, of which the symmetric block matrix form is a special case. In related work, Baxter [242], [247], Devinatz [249], Ibragimov [251], and others have studied the asymptotic behavior of the error in Wiener filtering for quite general classes of stationary processes. These general techniques can also be usefully adapted to Kalman filters.

To close this section, we shall point out a connection with innovations that can be useful in generalizing several of the previous results to certain classes of nonstationary processes. For this we recall the isometric mapping (64)–(66) between the space of random variables and the space of polynomials on the unit circle under which $y(k) \leftrightarrow z^k$. Now, given $\{y(0),y(1),\cdots\}$, it is natural, following Wold [206], to consider the orthogonalized innovations sequence, say,

$$\varepsilon(0) = y(0) = I\cdot y(0)$$

$$\varepsilon(1) = y(1) - \hat{y}(1 \mid 0) = y(0) + A_{1,1}y(0).$$

$$\vdots \qquad\qquad \vdots$$

By our mapping each of the innovations $\varepsilon(N)$ corresponds to a polynomial $A_N(z) = I + A_{N,N}z + \cdots + A_{N,1}z^N$, and these polynomials must be orthogonal. In other words, Szegö's orthogonal polynomials are the images of the innovations under the isometric mappings (65), (66). Many properties of orthogonal polynomials can be interpreted as properties of innovations and vice versa. For example,

Szegö's classic result that the polynomials $A_N(z)$ have all their roots inside the unit circle, so that $A_N^{-1}(z)$ can be regarded as the transfer function of a bounded filter, is the analog of the fact that the transformation between a discrete-time process $y(\cdot)$ and its innovations is boundedly reversible.

So far, we have only noted connections with known results. On the other hand, innovations can be defined for large classes of processes besides stationary discrete-time sequences. Therefore, we have the possibility of obtaining various generalizations of the previous results and in particular of the classical theory of orthogonal polynomials. (See also [248a].)

These possibilities provide additional motivation for going on to a deeper study of innovations and the Wold representation.

VIII. Canonical Representations of Continuous-Time Processes

The concept of innovations processes was introduced for stationary discrete-time processes via the Wold representation theorem. The continuous-time decomposition, which was obtained by different methods by Krein [215], Karhunen [222], and Hanner [221], is a natural generalization of the discrete-time representation. It contains a regular part plus a deterministic part, which will be absent if and only if condition (72) is met, as we shall henceforth assume for convenience of writing. Then the Wold representation is

$$y(t) = \int_{-\infty}^{t} g(t-u)\, dE(u) \tag{91}$$

where $E(\cdot)$ is a process with uncorrelated increments, and the integral is the so-called Wiener stochastic integral, cf. Doob [225, ch. IX]. There are many kernels $g(\cdot)$ that can be used in the previous representation, as we shall see later, but there is always a particular one $g_0(\cdot)$ such that $y(\cdot)$ and $E(\cdot)$ are *causally equivalent* in the sense that any finite-variance random variable linearly dependent on $\{E(u), u \le t\}$ can also be calculated by linear operations on $\{y(s), s \le t\}$ and vice versa. In this case, we shall say that the Wold representation is *canonical*, or that

$$y(t) = \int_{-\infty}^{t} g_0(t-u)\, dE(u) \tag{92}$$

is an *innovations representation* of $y(\cdot)$, and we shall call $E(\cdot)$ the *innovations process* of $y(\cdot)$. This can be rewritten in a form closer to Wold's (cf. [60a])

$$y(t) = \int_{-\infty}^{t} g_0(t-u)e(u)\, du, \qquad e(u) = dE(u)/du.$$

However, $e(\cdot)$ is now a continuous time "white noise" process, so that the differentiation of $E(\cdot)$ is not valid in a classical sense, but only in the sense of generalized functions. However, as long as one is at most concerned with a white noise process (and not any of its generalized derivatives), there is no need to introduce the machinery of generalized processes; it suffices to work with the integrated process of uncorrelated increments $E(\cdot)$ and the notion of Wiener stochastic integrals. [The situation is analogous to that in deterministic system theory, where we can use Heaviside functions and Stieltjes integrals to handle singularities that are at worst delta functions.] Even when the second approach is taken, the concept of white noise is still very convenient and useful, as engineers have long known. A thought-provoking example is provided by the development of general series expansions for random processes, which were known to engineers much before they were formally discovered by mathematicians (cf. Section X).

However, as the mathematical sophistication of the field increases, there is a tendency to uncritically reject the use of white noise. This I believe is a serious mistake, which delays for many potential users the appreciation of several general and useful results that may have been originally derived in a more abstract context. Moreover, at the very least it can be a powerful guide to our intuition and a hedge against many unfruitful investigations. It may be apt to note that, according to Doob [253a], it was Wiener who first "showed, and applied repeatedly, that the [process $E(\cdot)$] acts as though [its] derivative process exists, is stationary and has a constant spectral density." Doob's book [225, pp. 435–436, p. 533, pp. 546–547] has some nice examples of how white noise can be usefully employed even in a rigorous exposition of stochastic processes addressed to mathematicians. A recent book of Hida [295] stresses the importance of white noise and this has been further reinforced by recent work of Hida and of McKean [323]. Rao has developed perhaps the most general results to date [305], studying more white noise processes than those defined as the derivatives of processes with independent increments [225].

Properties of Canonical Kernels

Karhunen and others have studied the kernels $g(\cdot)$ and $g_0(\cdot)$ in some detail. Karhunen [222] showed that all $g(\cdot)$ that could serve to define the Wold representation (91) are of the form

$$g(t) = \int \exp(i\omega t) G(\omega) \frac{d\omega}{2\pi}$$

$$G(\omega) = \lim_{\sigma \to 0} G(s), \qquad s = \sigma + i\omega$$

$$G(s) = A(s)G_0(s) \tag{93}$$

where

$$G_0(s) = \exp\left[\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1 - i\lambda s}{s - i\lambda}\, \frac{\log \dot{F}(\lambda)}{1+\lambda^2}\, d\lambda\right]$$

$$\dot{F}(\lambda) = \text{the power spectral density of the process } y(\cdot) \tag{94}$$

and

$$A(s) = \text{transfer function of an "all-pass" system.}$$

Any such all-pass transfer function can be further decomposed into

$$A(s) = A_0 A_1(s) A_2(s) A_3(s) \tag{95}$$

where

A_0 = constant of magnitude 1 (called a trivial all-pass function)

$$A_1(s) = \prod_k \frac{s_k - s}{\bar{s}_k + s} \cdot \frac{|s_k - 1|}{s_k - 1} \cdot \frac{|s_k + 1|}{s_k + 1} \quad \text{(Blaschke product)}$$

$$\text{Re } s_k > 0, \qquad \Sigma \text{ Re } s_k/1 + |s_k|^2 < \infty \tag{96}$$

$$A_2(s) = e^{-\alpha s} \quad \text{(pure delay)} \tag{97}$$

$$A_3(s) = \exp\left[-\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1 - i\lambda s}{s - i\lambda}\, d\beta(\lambda)\right] \quad \text{(singular part)} \tag{98}$$

where $\beta(\cdot)$ is a nondecreasing function whose derivative vanishes almost everywhere, and $0 < \beta(\infty) - \beta(-\infty) < \infty$. The $\{A_i\}$ all have unit modulus along the $i\omega$ axis, so that

$$|G(i\omega)| = |G_0(i\omega)|. \tag{99}$$

Also, for all but A_0 we have

$$|A_i(s)| < 1, \qquad \text{Re } s > 0, \qquad i = 1,2,3. \tag{100}$$

If we restrict $G(s)$ to being rational, then $A_2(s)$ and $A_3(s)$ will not appear and the so-called Blaschke product in $A_1(s)$ will be finite ($k < \infty$). Note that all the poles of $A_1(s)$ must be in the left half-plane (corresponding to causality, since the region of definition of all our functions here includes the $i\omega$-axis), though there can be zeros in the right half-plane. These results were obtained by using classical results on bilateral Laplace transforms (cf. Paley and Wiener [204]) and the so-called Hardy functions (see Hoffman [244] and Duren [258] for recent accounts). It may be of interest that the decomposition for $G(s)$ was obtained before Karhunen and Krein by Krylov [262] in a purely mathematical study of the transforms of "one-sided" (causal) time functions. Such results have again become of interest in recent studies on infinite-dimensional realization theory [307].

Karhunen proved that

$$G(s) = A_0 G_0(s) \tag{101}$$

is a necessary and sufficient condition for

$$L_2(y; t) = L_2(E; t), \qquad -\infty \le t \le \infty \tag{102}$$

where

$L_2(y; t)$ = the linear space of all finite linear combinations of $\{y(s), s \le t\}$, and their limits in the covariance norm. (103)

For all other $G(s)$, we have

$$L_2(y; t) \subseteq L_2(E; t), \qquad -\infty \le t \le \infty.$$

This explains the names canonical representation and innovations representation for (92). We note that $G_0(s)$ and $A(s)$ are sometimes, following Beurling [264], [312], called "outer" and "inner" functions, respectively.

There are various other interesting properties of the canonical kernel $g_0(\cdot)$, some of which would seem to be useful in developing adaptive filters [38], [173], [55].

1) Let $f(\cdot)$ be a square-integrable function on $(-\infty,\infty)$. Then

$$\int_{-\infty}^{t} g_0(t - u)f(u)\, du = 0, \qquad t \le 0,$$

$$\text{if and only if } f(t) = 0, \qquad t \le 0. \tag{104}$$

This property was discovered by Karhunen [222].

2) The canonical kernel has maximum partial energy in the sense that

$$\int_0^t g_0^2(u)\, du \ge \int_0^t g^2(u)\, du, \qquad \text{all } t > 0 \tag{105}$$

and all causal or noncausal $g(\cdot)$ with $|G(i\omega)| = |G_0(i\omega)|$. This property was given in 1962 by Robinson [271], though closely related results were also noted by Lévy in the mid-fifties (see, e.g., [266, p. 140]). See also [317].

3) The set $\{g_0(t - r), r \ge 0\}$ spans $L_2(0,\infty)$. This last result is due to Beurling [264] and provides a version for L_2 of a famous theorem of Wiener's that the Banach space L_1 can be spanned by the translates of a function with positive Fourier transform.

4) Let $G(s)$ be rational and let

$$G(i\omega) = |G(i\omega)| \exp i\phi(i\omega). \tag{106}$$

Then $\phi(\cdot)$ is called the *phase lag* of the filter $G(\cdot)$ and $-d\phi(i\omega)/d\omega$ is called the *group delay*. Of all filters with the same "gain" $|G(i\omega)|$, $G_0(\cdot)$ has the smallest phase lag and group delay. This is often labeled the *minimum-phase property* of $g_0(\cdot)$ (see [263]).

Applications to Prediction

It should be clear (as noted by Krein and Karhunen) that the innovations representation (92) can be used to solve the prediction problem just as in the discrete-time case. In fact we have

$$\hat{x}(t + \alpha \mid t) = \int_{-\infty}^{t} g_0(t - u)\, dE(u), \qquad \alpha > 0. \tag{107}$$

Of course, the difficult thing is to explicitly calculate $\{E(u), u \le t\}$ from $\{x(u), u \le t\}$. For rational $G_0(\cdot)$ the solution is simple: just pass $x(\cdot)$ through the filter with transfer function $[G_0(s)]^{-1}$. (In principle, this is what we do in general, but care has to be taken in defining the inverse transfer function.) Thus we now have an alternative basically probabilistic approach to the continuous-time prediction problem, without bringing in any Wiener–Hopf equations.

These ideas were rediscovered by Bode and Shannon [5], and Zadeh and Ragazzini [6], though not in as much mathematical generality (only for rational spectra). These authors recognized very clearly that what was involved in the innovations approach was 1) replacing, causally and without loss of information, the given process by a "simpler" process, and 2) solving the estimation problem for the simpler process. So far we have used white noise as a simpler process, but it is not necessarily the only such process. For example, Zadeh and Ragazzini [6] show how we may use a process with spectral density equal to the numerator of the spectral density of the original process. This idea was rediscovered for discrete-time autoregressive-

moving average processes by Rissanen and Barbosa [131] (see also Whittle [32]) and has been exploited in several recent papers [175], [192], [184]. In principle, there are many other possibilities as well, and studies of "generalized innovations processes" have been made by Kailath and Duttweiler [311], Kallianpur and Oodaira [321], and Rozanov [314], [325].

Nonstationary Processes

The cited work up to 1950 was all for stationary processes and used frequency-domain methods. In 1950 Hanner [221] gave a purely time-domain derivation of the continuous-time Wold decomposition, and this raised the possibility of extensions to nonstationary processes. The first results in this direction were perhaps those of Yaglom and Pinsker [265], who studied nonstationary processes that had stationary increments of order n. The simplest example is the Wiener process, which has stationary increments of first order, and can be represented as the integral of white noise. The general case was first studied by Lévy [260], [274], who sought representations of the form

$$y(t) = \int_0^t g(t,u)\, dE(u), \qquad 0 \le t \le T \le \infty. \tag{108}$$

When

$$L_2(y;\, t) = L_2(E;\, t) \tag{109}$$

he called the representations *proper canonical,* using the word *canonical* for the case $L_2(y;\, t) \subseteq L_2(E;\, t)$. This terminology has now been abandoned. The covariance of a process as in (108) is

$$R(t,s) = \int_0^{t\wedge s} g(t,u)g(s,u)\, du, \qquad t\wedge s = \min(t,s). \tag{110}$$

Now Lévy asked whether, given a covariance $R(t,s)$, one could find a suitable function $g(t,u)$. He did not obtain a general solution to this problem, but among other results, he did obtain the following useful one.

Let

$$R(t,s) = t\wedge s + \int_0^t \int_0^s K(u,v)\, du\, dv, \qquad 0 \le t, s \le T \tag{111a}$$

$$= \int_0^t \int_0^s [\delta(u - v) + K(u,v)]\, du\, dv \tag{111b}$$

where $K(\cdot,\cdot)$ is a continuous symmetric function of two variables. Assume that the eigenvalues of $K(\cdot,\cdot)$ on the square $[0,T] \times [0,T]$ are greater than -1, or equivalently that $R(\cdot,\cdot)$ is strictly positive definite on this square. Next determine a function $h(\cdot,\cdot)$ as the unique solution of the Wiener–Hopf type of equation

$$h(t,s) + \int_0^t h(t,u)K(u,s)\, du = K(t,s), \qquad 0 \le s \le t \le T. \tag{112}$$

Let $k(\cdot,\cdot)$ be the so-called resolvent function of $h(\cdot,\cdot)$, defined by the equation

$$h(t,s) - k(t,s) = \int_0^t k(t,u)h(u,s)\, du. \tag{113}$$

Then the innovations representation is

$$y(t) = \int_0^t \left[1 + \int_u^t k(\tau,u)\, d\tau\right] dE(u). \tag{114}$$

This is a useful result, though it is a bit difficult to see the rationale for the steps (see some following comments by Hida). We shall give an explanation in Section IX.

Although Lévy [266] noted that his method applied to certain more general kernels, e.g., those that were an α-fold double integral of $[\delta(t - s) + K(t,s)]$ (cf. (111)) with α not necessarily integral, he was unable to prove that (110) always has a solution, or, equivalently, that an innovations representation always exists for a (nondeterministic) process. The reason for this failure appeared only a few years later, when in 1960 Cramér [268] and Hida [269] independently discovered a new dimension to this problem, finding that a single kernel $g(\cdot,\cdot)$ does not suffice in general. The proper Wold decomposition for a finite-variance nonstationary process is

$$y(t) = \sum_1^N \int_0^t g_i(t,u)\, dE_i(u) + \psi(t) \tag{115}$$

where $\psi(\cdot)$ is a deterministic process, and the $\{E_i(\cdot)\}$ are orthogonal-increment processes, uncorrelated with each other. The number N is uniquely determined by the covariance of $y(\cdot)$, though the $\{g_i\}$ and $\{E_i(\cdot)\}$ are not. N is called the *multiplicity* of the process $y(\cdot)$, and if $N > 1$ the representation (115) is called a generalized canonical representation. The multiplicity N can be infinite [273], even though all presently known examples of processes with $N > 1$ have rather pathological kernel functions $\{g_i(t,u)\}$. Hitsuda has very recently discovered [320] that a process of the form

$$w_1(t) + f(t)w_2(t), \qquad t \ge 0 \tag{116}$$

where $w_1(\cdot)$ and $w_2(\cdot)$ are independent Wiener processes, will have multiplicity 1 if $f(\cdot)$ is absolutely continuous with a square-integrable derivative. However, if the derivative is not square-integrable on every open interval $(1,m) \subset [0,\infty)$, then the multiplicity is 2. The multiplicity is also 2 if $f(\cdot)$ has unbounded variation everywhere.

Hida [269] developed some of Lévy's ideas more clearly and obtained several new results. In his words "[Lévy's] pioneering works contain some points difficult for us to follow. The main aim of this paper is to establish his theory systematically and to prove some new facts." Among other things, Hida proved in great generality that if a process has a representation of the form (108), then it always has a canonical (innovations) representation obeying (109). He also gave the following extension of the criterion (104) of Karhunen: the representation (108) is an innovations

representation if and only if

$$\int_0^t g(t,u)f(u)\,du = 0, \qquad \text{for every } t \leq T \quad (117a)$$

implies

$$f(u) = 0, \qquad \text{a.e. on } [0,T]. \quad (117b)$$

These results are for processes of multiplicity one and are difficult to extend to the general case (115), because of the nonuniqueness of the $\{g_i\}$ and $\{E_i\}$.

Processes of Multiplicity One

It is useful to identify processes whose multiplicity is 1, since statistical applications will be easier in this case.

As noted earlier, the results of Krein, Karhunen, and Hanner show that nondeterministic stationary processes have multiplicity one. Cramér [268], [273], following the ideas of Hanner, obtained several interesting results. For example, he showed that nondeterministic discrete-time processes, stationary or not, always have multiplicity one, which is in sharp contrast to what we have just noted for continuous-time nonstationary processes. This result shows that some care must be exercised in studying continuous-time problems via discretization. Some studies of this question have recently been made [318]. Some other special examples of processes of multiplicity one (e.g., wide-sense Markov processes and certain harmonizable processes) have been given by Cramér [275], [300], who was apparently unaware of the class of processes identified by Lévy [cf. (111)].

In the engineering literature, the great appeal of the innovations approach as described by Bode and Shannon and by Zadeh and Ragazzini led to various attempts to discover innovations representations for nonstationary processes. A survey of the results achieved up to 1961 is given in a paper by Zadeh [270] and a further review was given by Stear [94] in 1965. However, the results achieved until then did not seem particularly exciting. In 1966, Anderson [98] extended some of Stear's work to show that a canonical representation for covariances of the separable form (27) could be obtained if a certain Riccati equation could be solved. This was nice because such equations were familiar from the Kalman filter. However, the Riccati equation developed by Anderson was of a different form from that used in the Kalman theory, and therefore the nontrivial question of existence of a solution to this nonlinear equation had to be settled differently [122], [127]. Related results were obtained by Brandenburg and Meadows [134] and Geesey [281].

However, the work in [98], [134], [122], [127], proceeded in apparent ignorance of the fundamental results of Hida and Cramér. It is perhaps noteworthy that even Lévy's result, though it was described in Zadeh's 1961 survey paper, escaped attention until it was rediscovered in a different way by the present author in 1968. This will be described in the next section, where we shall show that the Lévy result and the occurrence of the Riccati equation in [98], [134], can be easily explained by means of some estimation results.

IX. Recent Results on Innovations Processes and Some Applications

The innovations process was derived in discrete time by use of the Gram–Schmidt technique, according to which vectors are successively orthogonalized by projecting each new vector on the manifold spanned by the previous vectors. For a random sequence $\{(y_i)\}$ this means forming

$$\varepsilon_i = y_i - \hat{y}_{i|i-1}.$$

The analogous procedure for a continuous-time process is to form

$$\varepsilon(t) = y(t) - \hat{y}(t \mid t-)$$

but this will be identically zero for any process with continuous paths. However, if

$$y(t) = z(t) + v(t), \qquad 0 \leq t \leq T$$

and

$$E\int_0^T |z(t)|\,dt < \infty \quad (118)$$

then we will have a nontrivial innovations process

$$\varepsilon(t) = y(t) - \hat{y}(t \mid t-) = y(t) - \hat{z}(t). \quad (119)$$

Notice that (118), (119) is the time-domain analog of the Wiener filter formula (16). The process $\varepsilon(\cdot)$ has already been encountered in the Kalman filter (cf. (24)), and in that context it was recognized by a number of people ([109], [111], [114]) that $\varepsilon(\cdot)$ is a white noise process with

$$E\varepsilon(t)\varepsilon'(s) = Ev(t)v'(s) = I_p\delta(t-s). \quad (120)$$

However, the proofs of (120) in all these references were based on knowledge of the Kalman formulas for $\hat{x}(\cdot)$ and the mean-square error $P(\cdot)$. It was apparently not suspected that (120) holds quite generally; in fact my belief in this was the object of much skepticism at first. However, in April 1967, I obtained a proof initially via the Wiener–Hopf equation (10), and then by a simple use of the fundamental property that $z(t) - \hat{z}(t) \perp \{y(\tau), \tau < t\}$. This proof is described in [116, appendix I] (see also [298] and [291]). Thus we now had the opportunity to use the innovations method for *nonstationary* continuous-time processes.

During 1966–1967, the author, his students, Duncan and Frost, and Clark were studying nonlinear filtering. In that context it had become clear by a theorem of Lévy's (Doob [225, p. 384]) that if (120) could be established, one could obtain the striking fact that if $v(\cdot)$ were Gaussian (but not necessarily $z(\cdot)$), and if $\hat{z}(\cdot)$ was the least-squares estimate (not necessarily linear) then $\varepsilon(\cdot)$ would be not only white, but also Gaussian! This general result thus followed from our proof in April 1967; Frost simultaneously established it [279] in a different way using the Itô differential rule. Shiryaev later informed the author that he had found a similar result in 1966 [101, p. 22] in trying to make the evolution equations for nonlinear filtering mathematically meaningful. Namely, one had stochastic integrals with respect to the process $\varepsilon(\cdot)$ and these would not make sense unless $\varepsilon(\cdot)$ was a martingale process. The same reason led

to an independent discovery by Kallianpur (personal communication, 1969). By now, several proofs of (119), (120) have been given under increasingly general assumptions, with the most recent results being in Kailath [304] and Meyer [324]. Specific extensions to discontinuous processes with independent increments have been made by Frost [294], [302], Snyder [172], Bremaud [308], and in some generality by Segall [326]. By now, the role of martingale theory in this question has become much clearer, e.g., the innovations theorem is just the Meyer–Doob decomposition for semimartingales with respect to the sigma fields of the observation process [326]. However, this is not the place for an exposition of this topic.

The first application of the previous result was to deduce a general formula for the likelihood ratio in terms of the nonlinear estimate $\hat{z}(\cdot)$ (cf. [291], [297]). Such formulas had been obtained by Schweppe [93] for Gaussian $z(\cdot)$ and, independently and in more generality, namely, for Markov $z(\cdot)$ by Stratonovich and Sosulin in a series of papers beginning in 1964 ([89], [102], [141]). However, the arguments of these papers were rather cumbersome and, at the author's suggestion and with his help, Duncan succeeded in streamlining and clarifying Stratonovich's proof (cf. [108], [115]).

A popular method for attacking such detection problems (and related nonlinear estimation problems) is via the use of a generalized Bayes' rule; the likelihood ratio is first written for a fixed sample of the signal process, and then one averages over all possible signal paths. (The function-space version of this familiar technique was first introduced into the engineering literature by Bucy [91], and it is sometimes called Bucy's representation theorem. Bucy has given proofs under various conditions (cf. [131], [136]), apparently unaware that the most general form appeared in a 1968 paper by Kallianpur and Striebel [118].) This was essentially the method used by Stratonovich and Sosulin and then clarified and extended by Duncan [108], [115], [293]. However, it has a serious limitation. If the signal process can depend upon past observations of signal plus noise, as happens in feedback communication and control problems, then clearly the conditioning-and-averaging procedure of the generalized Bayes' rule cannot be followed, because fixing the signal may also fix the observation. It seems that the only way to overcome this difficulty is to use the innovations process. This is done for additive Gaussian noise in [296], [297], with a heavy use of martingale theory. It may be of interest that the general likelihood ratio formula was conjectured in May 1967, some time before much of the machinery for a rigorous proof was made available through a now famous paper of Kunita and Watanabe [278], published in late 1967, and a related December 1968 paper [282] of Hitsuda.

Applications of the innovations process to linear problems actually came after the application to nonlinear detection problems and were partly stimulated by the paper of Schweppe [93], just cited. Let

$$\hat{z}(t \mid T) = \int_0^T H(t,s)y(s)\,ds$$

and

$$\hat{z}(t) = \int_0^t h(t,s)y(s)\,ds \tag{121}$$

and assume that the signal $z(\cdot)$ and the noise $v(\cdot)$ are completely uncorrelated. Then from Schweppe's paper one can deduce that

$$H(t,s) = h(t,s) + h(s,t) - \int_0^T h(\tau,t)h(\tau,s)\,d\tau. \tag{122}$$

It turned out that this relation could be proved directly by using an integral equation identity of Siegert, Krein, and Bellman [58], [228], [231]. A proof via innovations was desired, but for this it was necessary to show that the process $\varepsilon(\cdot)$ spans the same space as the observation process $y(\cdot)$. The author succeeded in doing this in November 1967 by the following argument [116], [165].

We can write the relation

$$\varepsilon(t) = y(t) - \hat{z}(t) = y(t) - \int_0^t h(t,\tau)y(\tau)\,d\tau \tag{123}$$

symbolically as

$$\varepsilon = (I - h)y. \tag{124}$$

Now we can find y from ε if $(I - h)^{-1}$ exists, a sufficient condition for which is (Smithies [386]) that h be Volterra and square-integrable on $[0,T] \times [0,T]$. Causality is equivalent to the Volterra property for such functions (but not necessarily otherwise (cf. [311]), and it is easy to see that the square-integrability will follow from the assumption that $z(\cdot)$ has finite expected energy

$$E\int_0^T z^2(t)\,dt < \infty. \tag{125}$$

Therefore in the linear case, we know that under quite general circumstances $\varepsilon(\cdot)$ is a true innovations process: for every t, $\{(\varepsilon(\tau), \tau < t)\}$ spans the same linear space of random variables as $\{(y(\tau), \tau < t)\}$. With this fact, the Kalman filter can be readily derived [116], [165], and several related linear problems solved [117], [181]. Unfortunately, the analogous result for the nonlinear problem has still not been proved, except for uniformly bounded $z(\cdot)$ [283] (see also [301, appendix I]). However Fujisaki, *et al.* [309] have elegantly circumvented this lack, by proving that $\hat{z}(\cdot)$ can still be written as a stochastic integral with respect to the innovations process, after which the remaining calculations are straightforward. Nevertheless, a proof of the equivalence would still have various benefits, since it can also be regarded as a very general existence result for solutions of functional stochastic equations. Moreover, there is a feeling, not unshared by the author, that the term "innovations process" should only be used for $\varepsilon(\cdot)$ when it can be proved to contain the same information as the original observation process, i.e., whenever $\{y(s), s < t\}$ and $\{\varepsilon(s), s < t\}$ generate the same sigma fields. Nevertheless, many important results, e.g., those on signal detection [297] and on nonlinear filtering [309]

can be proved without using this property. Therefore, the process $\varepsilon(\cdot)$ certainly deserves a special name, and we use *innovations* in the hope that the equivalence question will be settled (sooner or later) in some generality for non-Gaussian $z(\cdot)$. The term "residual" might be used, but this term is already used in statistics to denote (in discrete-time) the process $r_i = y_i - H\hat{x}_{i|i}$, while the innovations process is $\varepsilon_i = y_i - H\hat{x}_{i|i-1}$.

The name innovations was apparently first used for the discrete-time sequence $\{y_i - \hat{y}_{i|i-1}\}$ by Wiener, Masani, and Kallianpur in their studies in the mid-fifties (personal communication from Masani). Some studies of its properties were made in the unpublished report [267], which is partially elaborated in Masani's survey [255, pp. 92–93]. Masani [255] and Loéve [248, p. 386] have pointed out that Lévy used the centered sequence $\{y_i - \hat{y}_{i|i-1}\}$ for various problems, beginning in the mid thirties. As noted earlier, Wold was the first to use it for linear estimation. Wiener and Kallianpur attempted to establish conditions for non-Gaussian $\{y_i\}$ under which the sigma field of the innovations would be causally equivalent to the observations. Masani [255] mentions this as one important problem "bequeathed to posterity" by Wiener. He continues "Also left to us is the extension of this theory to the continuous-parameter case. Here the absence of an atomic time-unit makes the problem of defining nonlinear innovations extremely hard; obviously all we may expect are virtual or differential innovations". As noted earlier [304], [324], [326], martingale theory shows how to do this for semi-martingales (processes of bounded variation plus a martingale, e.g., a Wiener or centered Poisson process).

The proof of equivalence in the linear case immediately gives a factorization result for the covariance of the process y. Let us define a Volterra function k via

$$(I - h)^{-1} = I + k \tag{126a}$$

or equivalently via the Volterra equation

$$h + hk = k. \tag{126b}$$

Then

$$\begin{aligned} I + K = Eyy' &= (I - h)^{-1}E\varepsilon\varepsilon'(I - h')^{-1} \\ &= (I - h)^{-1}(I - h')^{-1} \\ &= (I + k)(I + k'). \end{aligned} \tag{127}$$

Equivalently, we have a canonical representation for $y(\cdot)$, viz.,

$$y = (I + k)\varepsilon. \tag{128}$$

This factorization, so natural now, took some time to be recognized. Shepp had given a noncausal factorization, but could not solve the problem of causal factorization [276, pp. 332]. This is now accomplished by (127). This result having been obtained, it became clear that for continuous K we had just rediscovered the Lévy factorization (c.f. (111)–(114) of Section VIII, where y there is actually the integral of the previous y). However, our result was somewhat more general; assumption (125) does not require $K(\cdot,\cdot)$ to be continuous, only that $K(t,t)$ be integrable. Then in May 1968, we happened to discover a book by Gohberg and Krein [280] that showed that the square-integrability of $K(\cdot,\cdot)$ was sufficient. Later Hitsuda [282] gave a proof of this fact using martingale theory (cf. the discussion in [296]). Combining these results with a decomposition formula of Shepp's [276, theorem 7] for differential processes gives innovations representations for smooth processes as well [285], [181].

Besides these generalizations, however, a significant aspect of our approach was its provision of a neat stochastic interpretation for Lévy's calculations, namely that $h(t,s)$ is a least-squares filter, so that we can interpret Lévy's process $\varepsilon(t)$ as $y(t) - \hat{z}(t) = y(t) - \hat{y}(t \mid t-)$. Moreover, this interpretation suggests that when $z(\cdot)$ is generated by a lumped system, then a Riccati equation can be used to compute $\hat{z}(\cdot)$ and therefore also to solve the covariance factorization problem. This fact explains (see [149], [181]) the occurrence of the Riccati equation in the previously cited work of Anderson and Moore [98], [127], [140], Brandenburg [134], and Geesey [281, ch. II]. Moreover, it shows that other algorithms for computing estimates can also be used when appropriate, e.g., algorithms of Chandrasekhar type (Section VI).

Briefly, suppose that we have a process $y(\cdot)$ with covariance function as in (32)

$$\begin{aligned} I\delta(t - s) + M(t)\Phi(t,s)N(s)1(t - s) \\ + N'(t)\Phi'(s,t)M'(s)1(s - t). \end{aligned} \tag{128a}$$

Suppose also that the process $y(\cdot)$ arises from some state-space model of the form

$$\begin{aligned} \dot{x}(t) &= F(t)x(t) + G(t)u(t), \qquad x(0) = x_0 \\ \dot{y}(t) &= H(t)x(t) + v(t) \end{aligned} \tag{129}$$

with the usual assumptions (19)–(21) on $u(\cdot)$, $v(\cdot)$, x_0. The least-squares filter (22)–(24) for this state-space model can be rewritten as

$$\begin{aligned} \dot{\hat{x}}(t) &= F(t)\hat{x}(t) + K(t)\varepsilon(t), \qquad \hat{x}(0) = 0 \\ y(t) &= H(t)\hat{x}(t) + \varepsilon(t) \end{aligned} \tag{130}$$

where the gain function $K(\cdot)$, which is defined by

$$K(t) = P(t)H'(t) + G(t)C(t), \qquad P(t) = E\tilde{x}(t)\tilde{x}'(t) \tag{131}$$

can be calculated via the Riccati equation (27) for $P(\cdot)$, or directly via the Chandrasekhar-type equations of Section VI if F, G, H are constant.

Now, since $\varepsilon(\cdot)$ is known to be white, (130) can be regarded as another causal model for $y(\cdot)$ driven by a single white noise. Also, it is easy to calculate $\varepsilon(\cdot)$ from $y(\cdot)$: replace ε by $y - H\hat{x}$ in the differential equation, calculate $\hat{x}$, and then form ε as $y - H\hat{x}$. Therefore, for a process with a *known* state model (129), (130) defines the innovations representation (IR). (This simple fact, widely known by now, was to our knowledge first pointed out in [116, appendix D] and explicitly restated in [284].)

Now the general studies of Hida and Cramér on IRs (cf. Section VIII) have shown that the IR of a process $y(\cdot)$ can depend only upon the covariance of $y(\cdot)$. Therefore, F, K, H in (129) *must be determinable* (up to state-space transformations) from the covariance of $y(\cdot)$, no matter what state model we initially assumed. To do this we can calculate the covariance of $y(\cdot)$ as given by the particular model (129) and compare with the given expression (128) to make the nonunique identifications

$$M(t) = H(t), \qquad F(t) = \frac{d\Phi(t,s)}{dt}\,\Phi^{-1}(t,s)$$

$$N(t) = \Pi(t)H'(t) + G(t)C(t).$$

We now use these relations to express the parameters (F,H,K) of the IR (130) in terms of (M,Φ,N). Let

$$\Sigma(t) \triangleq E\hat{x}(t)\hat{x}'(t).$$

Then by the orthogonality of $\hat{x}$ and $\tilde{x}$ we have

$$Ex(t)x'(t) \triangleq \Pi(t) = \Sigma(t) + P(t)$$

so that we can write $K(\cdot)$, as given by (131), as

$$\begin{aligned} K(t) &= \Pi(t)H'(t) + G(t)C(t) - \Sigma(t)H'(t) \\ &= N(t) - \Sigma(t)M'(t). \end{aligned} \tag{132}$$

Moreover, using (130) and the fact that $\varepsilon(\cdot)$ is white readily yields

$$\begin{aligned} \dot{\Sigma}(t) = F(t)\Sigma(t) + \Sigma(t)F'(t) + [N(t) - \Sigma(t)M'(t)] \\ \cdot [N(t) - \Sigma(t)M'(t)]', \qquad \Sigma(0) = 0. \end{aligned} \tag{133}$$

The stochastic interpretation of $\Sigma(\cdot)$ guarantees its existence, and we see that (132) and (133) enable us to calculate $K(\cdot)$ from knowledge only of $M(\cdot)$, $N(\cdot)$, and $\Phi(\cdot,\cdot)$. Therefore, (130), (132), and (133) determine the canonical or innovations factorization of the covariance (128). The state model (129) was only used here to motivate the development of the IR, as defined by (130), (132), (133). The result can be deduced just from the assumption that the covariance in (128) is strictly positive definite (cf. [285], [181, appendix III]). However, the present derivation does have the advantage that it clearly displays the intimate relationship between the filtering problem and the factorization problem. Thus we see that the state vector of the IR is $\hat{x}(\cdot)$, so that $H(t)\hat{x}(t) = M(t)\hat{x}(t)$ can also be computed directly from knowledge of the covariance function; this gives us a proof of the result presented in Section V. Other applications are described in [181], [285], [292], [310].

As one example, we shall show how to obtain some matrix spectral factorization algorithms.

Multivariate Spectral Factorizations

Consider a covariance function $R_y(\cdot)$ of the form (128a) where M and N are constant $p \times n$ and $n \times p$ matrices, respectively, and

$$\Phi(t,s) = \exp F(t-s), \qquad F = \text{a stability matrix.} \tag{134}$$

The power spectral density matrix is defined by the values for $s = i\omega$ of the function

$$\begin{aligned} S_y(s) &= \int_{-\infty}^{\infty} R(t)e^{-st}\,dt \\ &= I + M(sI - F)^{-1}N + N'(-sI - F')^{-1}M'. \end{aligned} \tag{135}$$

This is not generally the form in which the power spectral density will be given, but let us postpone this aspect for a while. The problem is to factor $S_y(s)$ as [cf. (14)]

$$S_y(s) = S_y^{+}(s)S_y^{+}(-s)$$

where $S_y^{+}(s)$ is the transfer-function matrix of a causal and causally invertible system with p inputs and p outputs.

Now, for finite time, we have such a system in the IR (130), (132), (133). However, when F is stable, it is not hard to prove (see, e.g., [126], [135]) that as $t \to \infty$, $\Sigma(t)$ in (133) will tend to a constant matrix $\bar{\Sigma}$ and hence that $K(t)$ will tend to a constant matrix $\bar{K}$

$$\bar{K} = N - \bar{\Sigma}M' \tag{136}$$

where $\bar{\Sigma}$ is the unique nonnegative-definite solution of the algebraic Riccati equation

$$0 = F\bar{\Sigma} + \bar{\Sigma}F' + [N - \bar{\Sigma}M'][N - \bar{\Sigma}M']'. \tag{137}$$

Then the innovations representation (130) has the transfer function

$$S_y^{+}(s) = I + M(sI - F)^{-1}\bar{K} \tag{138}$$

which we have designated $S_y^{+}(s)$ because in the limit it continues to provide the canonical (causal and causally convertible) factor of the power spectral density. From (135) and (138), we obtain the useful spectral factorization formula

$$\begin{aligned} I + M(sI - F)^{-1}N + N'(-sI - F')^{-1}M' \\ = [I + M(sI - F)^{-1}\bar{K}][I + \bar{K}'(-sI - F')^{-1}M']. \end{aligned} \tag{139}$$

We clearly have several procedures for computing $\bar{K}$ and thereby factoring $S_y(s)$. We can find the unique nonnegative-definite solution $\bar{\Sigma}$ of the nonlinear algebraic equation (137), or we can find $\bar{\Sigma}$ as the limiting solution of the Riccati differential equation (133); another more direct method is to find $\bar{K}$ as the limiting solution of the Chandrasekhar-type equations of Section VI. The differential equation procedures would be preferred because of their simplicity and automatic production of the right $\bar{\Sigma}$ or $\bar{K}$, but it is sometimes difficult to control the accumulation of computational errors until steady state is reached. However, the Chandrasekhar-type algorithms seem to behave quite well in this regard. The solution of the quadratic algebraic equation (137) involves choosing among the several possible solutions for $\bar{\Sigma}$ and is generally more laborious, though efficient eigenvalue–eigenvector methods have recently been proposed [78], [100], [160].

We now examine the problem of how to handle $S_y(s)$ that are not given in the form (135). The properties of power spectra show that we can easily write $S_y(s)$ in the form

$$S_y(s) = \frac{\text{a polynomial matrix}}{\Psi(s)\Psi(-s)} \tag{140}$$

where $\psi(s)\psi(-s)$ is a scalar polynomial that is the greatest common divisor of the denominators of all the entries in $S_y(s)$.

Now by a partial fraction expansion (or other means) we can decompose $S_y(s)$ as

$$S_y(s) = Z(s) + Z'(-s) \tag{141}$$

where $Z(s)$ contains all the terms corresponding to the left half-plane zeros of $\psi(s)\psi(-s)$. We can identify $Z(s)$ as the Laplace transform of the positive-time part of $R(t)$ [cf. (134), (135)]

$$Z(s) = \int_0^\infty [\tfrac{1}{2}I\delta(t) + Me^{Ft}N]e^{-st}\,dt \tag{142}$$

$$= \tfrac{1}{2}I + M(sI - F)^{-1}N. \tag{143}$$

Now $Z(s)$ can be regarded as the transfer function of a system with impulse response function $[\frac{1}{2}I + Me^{Ft}N]$, and therefore given $Z(s)$ we can find M, F, N by using one of several algorithms [129], [306], [360], going from rational transfer functions to state-variable realizations.

Thus we have some new methods for the classical multivariate spectral factorization problem. This problem has a long history and several different algorithms have been proposed by Wiener and Masani [12], Youla [26], Davis [28], Yaglom [23], Rozanov [22], Masani [35], Csaki and Fischer [36], Tuel [125], Strintzis [55], and others. The method based on solution of the nonlinear algebraic equation (137) was first derived in a different way by Anderson [104], who used certain connections, initially noted by Youla [26] and Kalman [75], [77], between spectral factorization and certain functions long familiar in network theory.

The point is that the $Z(s)$ in (142), (143) is not an arbitrary transfer function, but a "driving-point" impedance function, a property that Brune showed, in a 1931 dissertation that essentially founded network theory, is equivalent to $Z(s)$ being *positive real*, viz., that it obeys the conditions 1) all elements of $Z(s)$ are analytic in Re $s > 0$; 2) $Z(s)$ is real when s is real and positive; 3) $Z(s) + Z'(-s) \geq 0$, if Re $s > 0$. It can be shown (see, e.g. [177]) that equivalent conditions are that $Z(s) + Z'(-s)$ is a power spectral density matrix or that

$$Z(s) = \int_0^\infty R(t)e^{-st}\,dt$$

where $R(\cdot)$ is a nonnegative-definite function (i.e., a covariance function). These equivalences enable a considerable interplay between results in network theory, stability theory, control theory, and estimation (see, e.g., [77], [87], [158], [177]).

When $Z(s)$ has rational elements, an important tool in such studies has been the so-called *positive-real lemma*, first given by Yakubovic [73] and Kalman [75] for scalar $S_y(s)$, and extended to the matrix case by Popov [87] and Anderson (see references in [177]).

The positive-real lemma starts with a (nonunique) *minimal realization* of $Z(s)$ in the form

$$Z(s) = J + M(sI - F)^{-1}N \tag{144}$$

where minimality means that the square matrix F has lowest dimension among all possible F that could be used. Then it states that $Z(s)$ will be positive real if and only if there exists a real symmetric matrix $\overline{\Pi} \geq 0$ such that

$$\mathcal{M} = \begin{bmatrix} F\overline{\Pi} + \overline{\Pi}F' & N - \overline{\Pi}M' \\ (N - \overline{\Pi}M')' & J + J' \end{bmatrix} \geq 0. \tag{145}$$

Since the $\mathcal{M}$ matrix is nonnegative definite, it can be factored as

$$\mathcal{M} = \begin{bmatrix} L \\ W \end{bmatrix} [L' \quad W'] \tag{146}$$[7]

where the column size of L and W is arbitrary. Then an alternative statement is clearly that $Z(s)$ will be positive real if and only if there exist matrices L and W such that

$$F\overline{\Pi} + \overline{\Pi}F' = LL' \tag{147a}$$

$$N - \overline{\Pi}M' = LW \tag{147b}$$

$$J + J' = WW'. \tag{147c}$$

The significance of L and W is that they immediately give a factorization of $S_y(s)$. In fact, we can check that

$$S_y(s) = Z(s) + Z'(-s)$$
$$= [W + M(sI - F)^{-1}L][W' + L'(-sI - F')^{-1}M'. \tag{148}$$

There are many matrices $\overline{\Pi}$ that will satisfy (145), and consequently there will be many factorizations. The family of all such solutions has been studied by Anderson [177], Willems [158], Kucera [167], and Canabal [171]. The maximum and minimum $\overline{\Pi}$, when they exist, play a significant role in the analysis, with the minimum being the one that gives the innovations factorization; the maximum relates similarly to a certain dual system.

We cannot pursue such discussions any further here, but it may be interesting to note that related and in fact somewhat more general minimality properties were discovered by Krein in 1945 (cited in [22]) and by Masani [237]. The multivariate estimation problem has many fascinating aspects that are not generally known in the engineering literature, but we must content ourselves here to calling attention to the book [30] and two fine surveys by Masani [35], [45].

We conclude this section in a more engineering vein.

[7] Such factorizations of the discrete-time time-invariant generalization of the matrix in (145) are at the heart of the square-root algorithms mentioned briefly at the end of Section VI.

Transfer-Function Models Versus State Models

The many successes of the state model in recent years have led to an unwise neglect of more traditional methods and problems. Some examples were discussed in Sections VI and VII. Another example, briefly noted in Section IX, is the use of multivariate transfer-function models and frequency-domain analysis. Such models are essentially the only ones used in the statistical literature, where they are known as ARMA (autoregressive-moving average models). There are several interesting features associated with algorithms that work directly with such models, including the possibilities of fewer computations, easier proofs of stability, and more insight into certain structural aspects.

Briefly, suppose we have a process $y(\cdot)$ such that

$$y(n) + A_1 y(n-1) + \cdots + A_n y(0) = w(n)$$

where

$$w(n) = B_0 u(n) + B_1 u(n-1) + \cdots + B_m u(n-m)$$

and $u(\cdot)$ is a white-noise sequence. Then $w(\cdot)$ is a *moving-average* process. If $B_i = 0$, $i \geq 1$, then $y(\cdot)$ is an *auto-regressive* process; otherwise we have a *mixed* or *auto-regressive-moving average* process. The interesting point is that the innovations for $y(\cdot)$ are just the innovations for $w(\cdot)$

$$\hat{y}(n \mid n-1) = -A_1 y(n-1) - \cdots - A_n y(0) - \hat{w}(n \mid n-1)$$

so that

$$\varepsilon(n) = y(n) - \hat{y}(n \mid n-1) = w(n) - \hat{w}(n \mid n-1).$$

The process $w(\cdot)$ is often much simpler than $y(\cdot)$, e.g., m may be much less than n, or the $\{B_i\}$ may be constant while the $\{A_i\}$ are time variant or even nonlinear. These possibilities have already been exploited in [131], [175], [191], but more can be done. A useful stimulus to such researches is also provided by the relatively recent work of Popov [355], [359], Rosenbrock [356], Wang [358], Morf [191], Forney [361], Wolovich [363], and others, which has uncovered the close relationships between linear ARMA models and state-space models, thus enabling a fruitful combination of time- and frequency-domain methods.

X. Karhunen–Loève Expansions: Canonical Correlations and State Models

Those familiar with the textbooks on statistical communication theory in the last decade or so may be surprised that we have not referred to series expansions of random processes, and more particularly to the Karhunen–Loève (K–L) expansions (see, e.g., [341], [342], [40]). Consider a scalar Gaussian process $z(\cdot)$ with a continuous covariance $R_z(t,s)$ such that

$$\int_0^T \int_0^T R_z^2(t,s)\, dt\, ds < \infty.$$

The assumption of Gaussianness is made for terminological convenience; all statements can be translated in a standard way to apply to just "second-order" processes.

The K–L expansion of $z(\cdot)$ is

$$z(t) = \sum_1^\infty z_i \Psi_i(t) \tag{149}$$

where the $\{\Psi_i(\cdot)\}$ are eigenfunctions of $R_z(\cdot,\cdot)$, viz.,

$$\int_0^T R_z(t,s)\Psi_i(s)\, ds = \lambda_i \Psi_i(t), \qquad 0 \leq t \leq T$$

and

$$z_i = \int z(t)\Psi_i(t)\, dt, \qquad i = 1,2,\cdots.$$

It is known from the theory of integral equations that the $\{\Psi_i(\cdot)\}$ are orthonormal

$$\int_0^T \Psi_i(t)\Psi_j(t)\, dt = \delta_{ij}$$

and that (Mercer's formula)

$$R_z(t,s) = \sum_1^\infty \lambda_i \Psi_i(t)\Psi_i(s). \tag{150}$$

A simple calculation shows that the coefficients $\{z_i\}$ are uncorrelated

$$Ez_i z_j = \lambda_i \delta_{ij}.$$

If we temporarily write the random process as $z(t,\omega)$, ω being the probability-space variable, then the K–L expansion is a decomposition of a function of two variables into a sum of products of functions of one variable

$$z(t,\omega) = \sum_1^\infty z_i(\omega)\Psi_i(t).$$

Such decompositions are familiar from the partial differential equations of physics and their significance in random-process theory is basically the same. Since the $\{\Psi_i(\cdot)\}$ are deterministic, we can replace study of the uncountable family of random variables $z(t,\omega)$ by that of the countable family $\{z_i(\omega)\}$. The K–L expansions have the further useful property that the $\{z_i(\omega)\}$ are independent because $z(\cdot)$ is Gaussian, which simplifies many probabilistic calculations, e.g., determination of moments and convergence of sums. The K–L expansion (Karhunen [331], Loève [329]) was independently introduced by Kac and Siegert [330], [331], to simplify the calculation of the distribution of the output power from a nonlinear circuit (limiter-squarer-filter) driven by noise. On more abstract grounds, it had already been introduced in 1943 by Kosambi [328], an Indian statistician and Marxist philosopher, and also by Obukhov (in a 1946 dissertation, cited in [346]), Pugachev [339], and perhaps many others. The popularity of the K–L expansion (this terminology is now well entrenched) grew from its use in the 1950 Ph.D. dissertation [333] of Grenander to extend to stochastic processes the classical theories of statistical estimation and hypothesis testing, which had been developed for finite families of random variables.

The K–L expansion was soon used in estimation problems by Davis [335], Slepian [336], and Youla [337]. Its presentation in 1958 in the pioneering textbook of Davenport and Root [341] and the many applications in the widely used 1960 textbook of Helstrom [342] gave the K–L expansion a major place in the literature of the sixties.

Despite these many successes, however, the use of such expansions is diminishing for several reasons. One is of course that K–L expansions really apply only to Gaussian (or second-order) processes, while recently martingale theory has enabled significant headway to be made with non-Gaussian processes (see, e.g., the survey [327] by Wong). However, this deficiency does not apply to linear filtering, the main concern of our survey. Here a common complaint is that the K–L expansion does not lend itself to recursive calculation because the $\{z_i\}$ and $\{\Psi_i(\cdot)\}$ are not easily updated as T increases. Also, since the $\{z_i\}$ depend upon the values $z(t)$, for all $t \in [0,T]$, K–L expansions would seem to be more appropriate for smoothing problems (data over $[0,T]$) rather than causal filtering problems.

For example, consider the smoothing integral equation (9) for *uncorrelated* signal $z(\cdot)$ and noise $v(\cdot)$, say (in an obvious notation)

$$H(t,s) + \int_0^T H(t,\tau)R_z(\tau,s)\,d\tau = R_z(t,s), \qquad 0 \le t,s \le T. \tag{151}$$

The Mercer formula (150) for $R_z(t,s)$ now shows readily that the solution can be written

$$H(t,s) = \sum_1^\infty \frac{\lambda_i}{1+\lambda_i}\Psi_i(t)\Psi_i(s), \qquad 0 \le t,s \le T. \tag{152}$$

The computational value of such a solution is debatable because the $\{\lambda_i\}$ and $\{\Psi_i(\cdot)\}$ are difficult to compute, but at least its explicitness is often convenient. No similar solution appears possible for the filtering (or Wiener–Hopf) equation (10) because of the causality constraint $0 \le s < t \le T$. Nevertheless, recently a number of authors [352], [353], have shown that, with proper interpretations, series-expansion techniques can also be exploited in causal filtering problems. The key to these results can be found in an old device of Swerling's (cited in [155]). Swerling calculates the filtered estimate $\hat{z}(t \mid t)$ as

$$\hat{z}(t \mid t) = \sum_1^\infty \hat{z}_{i|t}\Psi_i(t)$$

where the $\{\hat{z}_{i|t}\}$ are *smoothed* estimates of the coefficients $\{z_i\}$ given data on $[0,t]$, and therefore can be determined by solving the smoothing equation(7). The filtered estimate can then be put together as an infinite combination of smoothed estimates. (This reflects in a different way our comment in Section II that (10) is a family (but not the obvious one) of equations of the form (9).) The recursive Kalman filter can be derived along these lines.

While one common criticism of the K–L approach can thus be partly met, there is another more serious difficulty. The ease of solving the smoothing equation (151) is heavily dependent upon the assumption that $z(\cdot)$ and $v(\cdot)$ are uncorrelated, which makes the right side equal to $R_z(t,s)$. Otherwise we would also have [cf. (9)] the term $Ez(t)v(s)$, and now there is no obvious solution. Some reflection will show that what really yielded the solution (152) of (151) was the *simultaneous expansions*

$$Ez(t)z(s) = R_z(t,s) = \sum_1^\infty \lambda_i\Psi_i(t)\Psi_i(s)$$

$$Ev(t)v(s) = \delta(t-s) = \sum_1^\infty 1\cdot\Psi_i(t)\Psi_i(s)$$

$$z(t) = \sum z_i\Psi_i(t)$$

$$Ez_iz_j = \lambda_i\delta_{ij}$$

$$Ev_iv_j = \delta_{ij}$$

and the fact that

$$Ez(t)v(s) \equiv 0 \Rightarrow Ez_iv_j = 0, \qquad \text{for all } i,j.$$

But how can dependence between $z(\cdot)$ and $v(\cdot)$ be reflected into the $\{z_i\}$ and the $\{v_i\}$, especially a one-sided dependence as in (4)? This question does not seem to have been raised in the engineering literature, rigorous or formal, even though dependence is needed to model many problems, e.g., when feedback is present.

The answer is that one should not treat $z(\cdot)$ and $v(\cdot)$ separately but should work with the observed process $y(\cdot) = z(\cdot) + v(\cdot)$, which has covariance

$$Ey(t)y(s) = \delta(t-s) + K(t,s).$$

$K(t,s)$ will not generally be a covariance when $z(\cdot)$ and $v(\cdot)$ are correlated, but since $Ey(t)y(s)$ is a covariance one can show that $K(t,s)$ (assumed to be continuous in t and s) has only a finite number of negative eigenvalues, and the Mercer formula extends to such functions as well (Riesz–Nagy [385, p. 242]). The emphasis on the observed process $y(\cdot)$, as against the signal and noise processes separately, is also the key to the development of the recursive Wiener filters; cf. the discussion of (128)–(133) in Section IX.

We have passed quickly over the above point that the white noise $v(\cdot)$ or its covariance $\delta(t)$ do not meet the conditions for the validity of the K–L expansion (149) or the Mercer formula (150). However, the validity of the expansion

$$\delta(t-s) = \sum_1^\infty \phi_i(t)\phi_i(s)$$

$\{\phi_i(\cdot)\}$ = any complete orthonormal family on $L_2[0,T]$

is usually argued on the grounds that any L_2-function $f(\cdot)$ can be correctly calculated as

$$f(t) = \int_0^T f(s)\delta(t-s)\,ds = \sum_1^\infty f_i\phi_i(t)$$

$$f_i = \int_0^T f(t)\phi_i(t)\,dt, \qquad i = 1,2,\cdots.$$

Therefore the $\{\phi_i(\cdot)\}$ can be chosen as, say, the eigenfunctions $\{\phi_i(\cdot)\}$ of some other covariance, say $R_z(\cdot,\cdot)$, and it is usually argued that we can write

$$v(t) = \sum_1^\infty v_i\Psi_i(t) + v_{rem}(t)$$

where the quantity $v_{rem}(\cdot)$ needed to make both sides "equal" is orthogonal to the signal process $z(\cdot)$ and can therefore be ignored. We do not wish to push this argument too far, though it has been quite successfully used in the literature. However, lest one get too scornful of what some people call "engineering nonsense," we shall show how it can be used to obtain (or let us say conjecture) a result that appeared only much later in the mathematical literature (not that this is a new phenomenon).

Expansions of a Wiener Process

White Gaussian noise can be thought of as the formal derivative of a Wiener process $w(\cdot)$ with continuous covariance function

$$Ew(t)w(s) = \min(t,s).$$

Now the eigenfunctions of this covariance are readily calculated

$$\Psi_n(t) = (2/T)^{1/2}\sin[(2n-1)\pi t/2T], \qquad n = 1,\cdots$$

so that the K–L expansion is

$$w(t) = \sum_1^\infty w_n\Psi_n(t)$$

where the $\{w_n\}$ are uncorrelated random variables with variances $\{4T^2/(2n-1)^2\pi^2\}$. However, since formally

$$w(t) = \int_0^t v(\tau)\,d\tau, \qquad v(\cdot) = \text{white noise}$$

we can also write

$$w(t) = \sum_1^\infty v_i\Phi_i(t)$$

where the $\{v_i\}$ are uncorrelated unit-variance random variables

$$\Phi_i(t) = \int_0^t \varphi_i(\tau)\,d\tau$$

and the $\{\varphi_i(\cdot)\}$ are *any* complete orthonormal set in $L_2[0,T]$.

Thus we can get many expansions for $w(\cdot)$, each with uncorrelated random variables $\{w_i\}$, but with *different* families of deterministic functions $\Phi_i(\cdot)$. The $\{w_i\}$ can be calculated as

$$w_i = \int_0^T \varphi_i(t)v(t)\,dt = \int_0^T \varphi_i(t)\,dw(t).$$

For most purposes these expansions are almost as useful as the K–L expansion (149). The previous result was first obtained by Shepp [276] and has since been extended to fairly general (non-Wiener) processes [350]. These general results can also be heuristically explained by using white noise, but we shall give a slightly different explanation.

Reproducing Kernel Hilbert Spaces

The time functions $\{\Psi_i(\cdot)\}$ in the K–L expansion are orthogonal over $[0,T]$ in the sense that

$$\langle\Psi_i,\Psi_j\rangle = \int_0^T \Psi_i(t)\Psi_j(t)\,dt = \delta_{ij}.$$

This property does not hold for the $\{\Phi_i\}$

$$\langle\Phi_i,\Phi_j\rangle \neq \delta_{ij}$$

but if we define a new inner product as

$$\langle a(\cdot),b(\cdot)\rangle_{H(w)} = \int_0^T \dot{a}(t)\dot{b}(t)\,dt + a(0)b(0)$$

then it is easy to see that

$$\langle\Phi_i,\Phi_j\rangle_{H(w)} = \delta_{ij}$$

for all choices of the $\{\Phi_i(\cdot)\}$. This inner product is appropriate for the Wiener process $w(\cdot)$. For other processes, we can determine suitable inner products that will make the corresponding expansions have simultaneously orthogonal random variables and time functions. The only special feature of the K–L expansion is that the time functions are orthogonal with respect to the $L_2(dt)$ inner product. However, there is nothing sacred about L_2 or about Lebesgue measure dt; we could, for example, use a measure $p(t)dt$, where $p(\cdot)$ is a weight function such that

$$\int_I\int_I R^2(t,s)p(t)p(s)\,dt\,ds < \infty.$$

This leads to expansions that are orthogonal in $L_2(p(t)dt)$ norm. In fact, to avoid dependence on such arbitrary weight functions $p(\cdot)$ or on the arbitrary family $\{\varphi_i(\cdot)\}$, it is desirable to try to seek an "intrinsic" norm associated just with the covariance of the process $w(\cdot)$. The norm $\langle\ ,\ \rangle_{H(w)}$ is just such a norm, and it is called a *reproducing kernel norm* because of the property

$$\langle R_w(t,s),m(s)\rangle_{H(w)} = m(t)$$

for all $m(\cdot)$, such that

$$\|m\|^2_{H(w)} = \langle m(\cdot),m(\cdot)\rangle_{H(w)} < \infty.$$

Such reproducing kernel Hilbert spaces (RKHS) were introduced into stochastic process theory by Loève in 1948 (cf. [274, appendix I]), and their usefulness in statistical applications has been made clear, notably by Parzen [347]. Many other references as well as some tutorial explanations and applications are discussed in [347], [349], [311]. However, it should be noted here that engineers have generally tended to think only of the space L_2 whenever Hilbert space is mentioned, perhaps on the grounds that L_2 is isomorphic to any Hilbert space. But the norm is the most important feature of a Hilbert space, and *isometries* (norm-preserving isomorphisms) are more significant than

additive isomorphisms. The theory of RKHS shows that there are many other quite different and quite useful Hilbert spaces besides L_2.

There are many other aspects of series expansions that could be discussed (some are noted in Wong's survey [327]), but we shall conclude with a discussion of a somewhat different type of process representation.

Canonical Correlations and State Models

Since series expansions generally contain an infinite number of terms, they cannot be used directly without truncation to a finite number, and this brings up the question of which terms to keep. It is usually suggested that we keep the coefficients that have maximum variance. However, calculations of least-squares estimates on this basis have not been very satisfactory, especially for processes with rational spectral density. The reason is *roughly* that the whole data interval enters equally into the determination of all the series coefficients, whereas, for example in prediction, the most recent observations should make a larger contribution. The state-space description of random processes reflects this circumstance better, but there is another quite general statistical technique that is roughly equivalent. This is the so-called theory of *canonical correlations*, which was independently developed in the mid-thirties by Hotelling and Obukhov (cf. [346]). [Incidentally, Hotelling was apparently the first to use eigenvalue-eigenvector decompositions (the finite analogs of the K–L expansion) in statistics (in a 1933 paper in the *Journal of Educational Psychology*).]

These authors proposed the following method for studying the interrelations between two (full-rank) families of random variables $\{X_1,X_2,\cdots,X_n\}$ and $\{Y_1,Y_2,\cdots Y_m\}$. First find the linear combinations

$$U_1 = \sum_1^n \alpha_{1i}X_i, \qquad V_1 = \sum_1^m \beta_{1i}Y_i$$

that have the largest cross-correlation coefficient

$$\rho_1 = EU_1V_1/\sqrt{EU_1^2EV_1^2}.$$

Next find linear combinations U_2 and V_2 that are uncorrelated with (U_1,V_1) and have maximum correlation coefficient, and so on until a new set of random variables $(U_1,U_2,\cdots,U_n, V_1,\cdots,V_m)$ has been found that span the same space as the $\{X_1,X_2,\cdots,X_n, Y_1,\cdots,Y_m\}$ and are pairwise uncorrelated, except for the pairs (U_i,V_i), $i = 1,\cdots,$ min (n,m). The actual calculation can be shown to be equivalent to the solution of a certain eigenvalue problem (see, e.g., [340]).

It is reasonable that in making inferences about the $\{Y_i\}$ from the $\{X_i\}$, the first few canonical variables $\{U_i\}$ should give us more information than the first few coefficients of the discrete-time analog of the K–L expansion. This observation was made by Yaglom [346], who with Gelfand generalized the theory to the case of continuous-time random processes $\{x(s), s \in S\}$ and $\{y(t), t \in T\}$, and used it for the calculation of mutual information [338]. Some especially interesting results are obtained when the two families of random variables are the past and future of the same random process. Yaglom has shown that the past and future (and in fact any two disjoint segments) of a nondeterministic continuous-time process have a finite number of canonical variables if and only if the process has a rational spectral density.

Recent studies in automata theory and algebraic system theory [119], [129], have shown that the analysis of a system in terms of past and future leads naturally to a "state-space" description. Related ideas can be recognized in an interesting but somewhat obscure paper by Levinson and McKean [345]. For a stationary process $y(\cdot)$, they introduce, among others, the subspaces

$B_{(a,b)}$ = the linear (Hilbert) space spanned by the random variables $\{y(\tau), a \le \tau \le b\}$

$B_+ = B_{(0,\infty)}$ = the future

$B_- = B_{(-\infty,0)}$ = the past

$B_{+|-}$ = the projection of B_+ on B_-

$$B_{0+} = \bigcap_{\delta>0} B_{(0,\delta)}.$$

B_{0+} is called the *germ field* [343] and $B_{+|-}$ is the *minimal splitting field* of past and future, viz., it is the smallest field such that, given $B_{+|-}$, B_+ is independent of B_-. [There are clearly many splitting fields, e.g., B_-; the proof that $B_{+|-}$ is the minimal field was given by McKean in a fascinating paper [344] on multidimensional Brownian motion.] These concepts are clearly related to the notion of state, and with this in mind we should not be surprised that $B_{+|-}$ is finite-dimensional if and only if $y(\cdot)$ has a rational spectral density or that $B_{+|-} = B_{0+}$ if and only if the spectral density has no zeros [345], [261a]. Recall that the state of a system can be determined from the output and its derivatives, without knowledge of the input, if and only if the transfer function has no zeros.

It can be proved that the canonical variables for the sets $\{y(\tau), \tau \ge 0\}$ and $\{y(\tau), \tau \le 0\}$ are a useful basis for the state-space $B_{+|-}$, and this fact has recently been cleverly exploited by Akaike [316] to study state-space modeling. In particular he obtains a stochastic interpretation of the Ho–Youla–Silverman and other algorithms for determining a minimal state-space realization of an impulse response. We cannot pursue these matters any further here, though as a final comment we may note that the canonical correlations would seem to be useful in more problems than those to which they have been explicitly applied so far.

XI. Concluding Remarks

In this survey, we have described several well established and widely used results and on occasion we have also

indicated some areas for further work and some new directions.[8]

One important trend is the growing interplay between linear system theory and linear filtering theory. This is reinforced by the growing realization in both fields of the importance of understanding the underlying structural features and invariants of dynamical linear systems and of the stochastic processes that can be generated from them. This structural knowledge is bound to be useful in all applications where linear systems and stochastic processes arise, ranging from long-distance communication problems to the analysis of single circuits.

Another obvious direction is into nonlinear filtering. Actually the last decade has seen a considerable effort to extend the Kalman filter to nonlinear problems. This vast area will have to be surveyed separately, but some important aspects should be mentioned here.

Recursive formulas for updating the least-squares estimate (the conditional mean) were first obtained by Stratonovich [65], [66], and Kushner [86]. However, it was found that in general the formulas involve all the conditional moments, so that an infinite set of simultaneous equations (or equivalently, a *partial* differential equation for the conditional probability density or the conditional characteristic functional) is necessary. There is still no consensus as to a satisfactory way of "truncating" this set of equations, which incidentally is also encountered in other fields such as fluid mechanics and quantum mechanics [388]–[390]. Furthermore, no other computationally satisfactory approaches to solving directly, even approximately, the partial differential equations seem to be at hand, though spline function approaches do hold out some promise.

At the moment, one of the chief benefits of having attacked the nonlinear problem was that it brought to the fore certain difficulties associated with the proper definition of the differential and integral equations used to describe nonlinear operations on white noise. For nearly the first time, as far as engineers were concerned, the use of different definitions of integrals made a difference to the answer, rather than just to the "rigor" of the proofs. It also affected the proper modeling of physical nonlinear problems. Wong and Zakai [277], McShane [277a], [277b], and others have made important contributions to this subject, but there are still many unresolved questions. Nevertheless, this work pointed the way to the introduction of martingale theory into communication and control problems, a fact whose significance will probably far outreach any specific nonlinear filtering problems. This point has been discussed at greater length in Wong's survey [327]; see also [298].

[8] It perhaps goes without saying that we have confined ourselves for many reasons to the *theory* of linear filtering and in particular to the *probabilistic* theory, where knowledge of all statistical parameters (e.g., means and covariances) is assumed. The development of a *statistical* theory of filtering will introduce several new dimensions, although it might be noted that the lack of a complete statistical theory does not seem to have significantly limited the successful use of the ideas of the probabilistic theory.

It should be noted that the previously mentioned unexpected difficulties with models and integrals lent the nonlinear recursions more than academic interest. Wonham gave the first rigorous proof, but only for signals taking finitely many values (e.g., finite Markov chains) [97]. Other proofs have since been constructed for more general signals, but under rather complicated and physically obscure conditions. In particular, dependence of the signal on feedback observations was excluded. As noted in Section IX, Fujisaki *et al.* have recently used the innovations approach to overcome most of these difficulties.

However, although such rigorous proofs are of interest, their main contribution is to highlight the fundamental difficulties in the way of practical optimum filtering. Despite major efforts (see, e.g., the proceedings of many recent symposia) the field is in some disarray. A good account of some of the more successful efforts is given by Jazwinski [138]. I believe that the situation is somewhat analogous to that of linear filtering in the mid-fifties, when the field was rapidly grinding to a halt amid a welter of numerous attempts at direct extensions of the Wiener filter. The Kalman filter provided a new impulse that moved things out of the doldrums into a new and fruitful direction. Similarly in nonlinear filtering it may be that attempts to solve the nonlinear filtering problem along the lines of the successful Kalman linear filter are misdirected. Some new approach needs to be uncovered.

Perhaps the way to begin is by lowering our sights by restricting ourselves to parameter estimation rather than to estimation of rather general stochastic processes. Even this is a difficult subject, which will not even be outlined here. However, I bring it up because recently information-theoretic[9] ideas have been found to be useful in getting error bounds for such finite-parameter problems (see, e.g., [377]–[381], [384]), and to a small extent for certain infinite-parameter (or stochastic-process) problems as well, see [383]. Furthermore, recently Blahut [382] and others have begun to show how the basic results of information theory can also be illuminated by the use of some simple hypothesis-testing and parameter-estimation problems. This interchange will be valuable and is in fact somewhat overdue, which brings me to my final point.

At many times in the twenty-five years of information theory there has been a not inconsiderable dissatisfaction with the scope, development, and application of the theory. This may or may not (probably not) have been justified, but it is interesting, at least in my opinion, that the fields of signal detection, estimation, and stochastic processes have not experienced such traumas. It seems to me that the reason lies in the actively pursued connections between these subjects and many other topics. "Strict-sense" information theory, although a beautiful and important subject, has suffered by its partially deliberate insularity and isolation. The numerous interconnections of statistical

[9] In the Shannon sense.

signal processing with other fields, as richly displayed even in the small subdomain of linear filtering that we have surveyed, is perhaps the surest guarantee of its continued vitality. It has never been other than a pleasure for me to have worked in such a field.

Acknowledgment

The last sentence really says it all: I am grateful to all who have contributed to this field. Several friends have generously provided comments on drafts of this paper.

Bibliography

A. Wiener Filtering and Related Topics

[1] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications.* New York: Technology Press and Wiley, 1949. (Originally issued in February 1942, as a classified Nat. Defense Res. Council Rep.)

[2] N. Wiener and E. Hopf, "On a class of singular integral equations," *Proc. Prussian Acad., Math.-Phys. Ser.*, p. 696, 1931.

[3] R. B. Blackman, H. W. Bode, and C. E. Shannon, "Data smoothing and prediction in fire-control systems," Research & Development Board, Washington, D.C., Aug. 1944.

[4] N. Levinson, "A heuristic exposition of Wiener's mathematical theory of prediction and filtering," *J. Math. Phys.*, vol. 25, pp. 110–119, July 1947; reprinted as an Appendix in [1].

[5] H. W. Bode and C. E. Shannon, "A simplified derivation of linear least square smoothing and prediction theory," *Proc. IRE*, vol. 38, pp. 417–425, Apr. 1950.

[6] L. A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of prediction," *J. Appl. Phys.*, vol. 21, pp. 645–655, July 1950.

[7] C. L. Dolph and H. A. Woodbury, "On the relation between Green's functions and covariances of certain stochastic processes and its applications to unbiased linear prediction," *Trans. Amer. Math. Soc.*, vol. 72, pp. 519–550, 1952.

[8] G. F. Franklin, "The optimum synthesis of sampled-data systems," Electron. Res. Lab., Columbia Univ., New York, Tech. Rep. T-6/B, May 1955.

[9] R. Jaffe and E. Rechtin, "Design and performance of phase-lock circuits capable of near-optimum performance over a wide range of input signal and noise levels," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 66–76, Mar. 1955.

[9a] P. Elias, "Predictive coding," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, Mar. 1955.

[9b] R. Price, "On entropy equivalence in the time- and frequency-domains," *Proc. IRE*, vol. 43, p. 484, Apr. 1955.

[10] A. M. Yaglom, *Theory of Stationary Random Functions*, translated from the Russian by R. A. Silverman. Englewood Cliffs, N.J.: Prentice-Hall, 1962. (Originally published as a survey paper in 1955.)

[11] M. C. Yovits and J. L. Jackson, "Linear filter optimization with game theory considerations," in *IRE Nat. Conv. Rec.*, pt. 4, pp. 193–199, 1955.

[12] N. Wiener and P. Masani, "The prediction theory of multivariate stochastic processes, Pt. I," *Acta Math.*, vol. 98, pp. 111–150, 1957; Pt. II, *ibid.*, vol. 99, pp. 93–137, 1958.

[13] G. C. Newton, L. A. Gould, and J. F. Kaiser, *Analytical Design of Linear Feedback Controls.* New York: Wiley, 1957.

[14] M. Shinbrot, "A generalization of a method for the solution of the integral equation arising in optimization of time-varying linear systems with nonstationary inputs," *IRE Trans. Inform. Theory*, vol. IT-3, pp. 220–225, Dec. 1957.

[15] F. J. Beutler, "Prediction and filtering for random parameter systems," *IRE Trans. Inform. Theory*, vol. IT-4, pp. 166–171, Dec. 1958.

[16] H. Laning and R. Battin, *Random Processes in Automatic Control.* New York: McGraw-Hill, 1958.

[17] S. Darlington, "Nonstationary smoothing and prediction using network theory concepts," *IRE Trans. Inform. Theory* (Special Suppl.), vol. IT-5, pp. 1–14, May 1959.

[18] P. Leonov, "On an approximate method for synthesizing optimal linear systems for separating signals from noise," *Automat. Remote Contr.*, vol. 20, pp. 1039–1048, 1959.

[19] A. V. Balakrishnan, "On a characterization of processes for which the optimal mean-square systems are of specified form," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 490–500, Sept. 1960.

[20] Y. W. Lee, *Statistical Theory of Communication.* New York: Wiley, 1960.

[21] V. S. Pugachev, *Theory of Random Functions and Its Applications in Automatic Control.* Moscow: Goztekhizdat, 1960; 2nd ed., 1964 (Transl.: Reading, Mass.: Addison-Wesley, 1966).

[22] Yu. A. Rozanov, "Spectral properties of multivariate stationary processes and boundary properties of analytic matrices," *Theory Prob. Appl.* (USSR), vol. 5, pp. 362–376, 1960.

[23] A. M. Yaglom, "Effective solutions of linear approximation problems for multivariate stationary processes with a rational spectrum," *Theory Prob. Appl.* (USSR), vol. 5, pp. 239–264, 1960.

[24] S. S. L. Chang, *Synthesis of Optimum Control Systems.* New York: McGraw-Hill, 1961.

[25] E. L. Peterson, *Statistical Analysis and Optimization of Systems.* New York: Wiley, 1961.

[26] D. C. Youla, "On the factorization of rational matrices," *IRE Trans. Inform. Theory*, vol. 7, pp. 172–189, July 1961.

[27] A. V. Balakrishnan, "An operator-theoretic formulation of a class of control problems and a steepest descent method of solution," *SIAM J. Contr.*, vol. 1, pp. 109–127, 1963.

[27a] H. C. Hsieh and R. A. Nesbit, "Functional analysis and its applications to mean-square estimation problems," in *Modern Control Systems Theory*, C. Leondes, Ed. New York: McGraw-Hill, 1965, pp. 97–120.

[28] M. C. Davis, "Factoring the spectral matrix," *IEEE Trans. Automat. Contr.*, vol. AC-8, pp. 296–305, Oct. 1963.

[29] E. Parzen, "A new approach to the synthesis of optimal smoothing and prediction systems," in *Mathematical Optimization Techniques*, R. Bellman, Ed. Berkeley, Calif.: Univ. California Press, 1963, pp. 75–108.

[30] Yu. A. Rozanov, *Stationary Random Processes.* Moscow: Fizmatgiz, 1963 (Transl.: A. Feinstein, San Francisco: Holden-Day, 1967).

[31] V. F. Pisarenko and Yu. A. Rozanov, "On some problems for stationary processes reducing to equations of the Wiener–Hopf type," *Probl. Pered. Inform.* (in Russian), vol. 14, pp. 113–135, 1963.

[32] P. Whittle, *Prediction and Regulation.* New York: Van Nostrand Reinhold, 1963.

[33] C. W. Helstrom, "Solution of the detection integral equation for stationary filtered white noise," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 335–339, July 1965.

[34] K. Steiglitz, "The equivalence of digital and analog signal processing," *Inform. Contr.*, vol. 8, pp. 455–467, 1965.

[35] P. Masani, "Recent trends in multivariate prediction theory," in *Multivariate Analysis*, P. R. Krishnaiah, Ed. New York: Academic Press, 1966.

[36] F. Csaki and P. Fischer, "On the spectrum factorization," *Acta Tech. Acad. Sci. Hung.*, vol. 58, pp. 145–168, 1967.

[37] S. G. Mikhlin and K. L. Smolitsky, *Approximate Methods for Solution of Differential and Integral Equations.* New York: American Elsevier, 1967.

[38] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. Goode, "Adaptive antenna systems," *Proc. IEEE*, vol. 55, pp. 2143–2159, Dec. 1967.

[39] J. F. Claerbout, "A summary, by illustrations, of least squares filters with constraints," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 269–272, Mar. 1968.

[40] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York: Wiley, 1968. Reviewed by R. A. Scholtz, *IEEE Trans. Inform. Theory* (Book Rev.), vol. IT-14, pp. 612–613, July 1968.

[41] I. F. Blake, "Linear filtering and piecewise linear correlation functions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 345–349, May 1969.

[42] W. M. Brown and R. B. Crane, "Conjugate linear filtering," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 462–465, July 1969.

[43] D. Slepian and T. T. Kadota, "Four integral equations of detection theory," *SIAM J. Appl. Math.*, vol. 17, pp. 1102–1117, 1969.

[44] D. L. Snyder, *The State-Variable Approach to Continuous Estimation, with Applications to Analog Communication Theory.* Cambridge, Mass.: M.I.T. Press, 1969. Reviewed by E. C. Posner, *IEEE Trans. Inform. Theory* (Book Rev.), vol. IT-18, p. 314, Mar. 1972.

[45] P. Masani, "Review of *Stationary Random Processes*, by Yu. A. Rozanov," *Ann. Math. Statist.*, vol. 42, pp. 1463–1467, 1971.

[46] J. J. Stiffler, *Theory of Synchronous Communications.* Englewood Cliffs, N.J.: Prentice-Hall, 1971. Reviewed by R. A. Scholtz, *IEEE Trans. Inform. Theory* (Book Rev.), vol. IT-18, pp. 218–219, Jan. 1972.

[47] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part II—Nonlinear Modulation Theory.* New York: Wiley, 1971. Reviewed by J. B. Thomas, *IEEE Trans. Inform. Theory* (Book Rev.), vol. IT-18, pp. 450–451, May 1972.

[48] K. Yao, "On the direct calculations of MMSE of linear realizable estimator by Toeplitz form method," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-17, pp. 95–97, Jan. 1971.

[49] ——, "An alternative approach to the linear causal least-square filtering theory," *IEEE Trans. Inform. Theory*, vol. IT-17, pp.

232–240, May 1971.
[50] Yu. A. Rozanov, "Some approximation problems in the theory of stationary processes," *J. Multivariable Anal.*, vol. 2, pp. 135–144, June 1972.
[51] J. A. Cochran, *Analysis of Linear Integral Equations.* New York: McGraw-Hill, 1972.
[52] W. C. Lindsey, *Synchronization Systems in Communication and Control.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.
[53] L. Prouza, "On generalized linear discrete inversion filters," *Kybernetika*, vol. 8, pp. 264–267, 1972; also, *ibid.*, vol. 6, pp. 225–240, 1970.
[54] J. Snyders, "Error expressions for optimal linear filtering of stationary processes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 574–582, Sept. 1972.
[54a] J. Snyders, "Error formulae for optimal linear filtering, prediction and interpolation of stationary time series," *Ann. Math. Statist.*, vol. 43, pp. 1935–1943, 1972.
[54b] M. G. Strintzis, "A solution to the matrix factorization problem," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 225–232, Mar. 1972.
[55a] D. G. Messerschmitt, "A geometric theory of intersymbol interference, Part I: Zero-forcing and decision-feedback equalization," *Bell Syst. Tech. J.*, vol. 52, pp. 1483–1519, 1973.
[55b] R. W. Lucky, "A survey of the communication theory literature: 1968–1973," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 725–739, Nov. 1973.
[55c] J. Salz, "Optimum mean-square decision feedback equalization," *Bell Syst. Tech. J.*, 1974.

B. Recursive Wiener and Kalman Filtering

[56] A. G. Carlton and J. W. Follin, Jr., "Recent developments in fixed and adaptive filtering," NATO Advanced Group for Aerospace R&D, AGARDograph 21, 1956.
[57] J. E. Hanson, "Some notes on the application of the calculus of variations to smoothing for finite time, etc.," Appl. Phys. Lab., Johns Hopkins Univ., Baltimore, Md., Internal Memo BBD-346, 1957.
[58] A. J. F. Siegert, "A systematic approach to a class of problems in the theory of noise and other random phenomena, Pt. 11," *IRE Trans. Inform. Theory*, vol. IT-13, pp. 38–43, Mar. 1957; Pt. III, *ibid.*, vol. IT-4, pp. 4–14, Mar. 1958.
[59] M. Blum, "Recursion formulas for growing memory digital filters," *IRE Trans. Inform. Theory*, vol. IT-4, pp. 24–30, Mar. 1958.
[60] B. Friedland, "Least-squares filtering and prediction of nonstationary sampled data," *Inform. Contr.*, vol. 1, pp. 297–313, 1958.
[61] P. Swerling, "First-order error propagation in a stagewise smoothing procedure for satellite observations," *J. Astronaut. Sci.*, vol. 6, pp. 46–52, Autumn 1959; see also "A proposed stagewise differential correction procedure for satellite tracking and prediction," RAND Corp. Rep. P-1292, Jan. 1958.
[62] R. E. Kalman, "On the general theory of control," in *Proc. 1st IFAC Cong.* London: Butterworth, 1960.
[63] ——, "Contributions to the theory of optimal control," *Bol. Soc. Mat. Mex.*, vol. 5, pp. 102–119, 1960.
[64] ——, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, pp. 34–45, Mar. 1960.
[65] R. L. Stratonovich, "Application of the theory of Markov processes for optimum filtration of signals," *Radio Eng. Electron. Phys.* (USSR), vol. 1, pp. 1–19, Nov. 1960.
[66] ——, "Conditional Markov process theory," *Theory Prob. Appl.* (USSR), vol. 5, pp. 156–178, 1960.
[67] M. Blum, "A stagewise parameter estimation procedure for correlated data," *Numer. Math.*, vol. 3, pp. 202–208, 1961.
[68] R. E. Kalman, "New methods of Wiener filtering theory," in *Proc. 1st Symp. Engineering Applications of Random Function Theory and Probability*, J. L. Bogdanoff and F. Kozin, Eds. New York: Wiley, 1963, pp. 270–388; also, RIAS, Baltimore, Md., Tech. Rep. 61-1, 1961.
[69] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Trans. ASME, Ser. D, J. Basic Eng.*, vol. 83, pp. 95–107, Dec. 1961.
[70] R. H. Battin, "A statistical optimizing navigation procedure for space flight," *J. Amer. Rocket Soc.*, vol. 32, pp. 1681–1692, 1962.
[71] J. D. McLean, S. F. Schmidt, and L. A. McGee, "Optimal filtering and linear prediction applied to a midcourse navigation system for the circumlunar mission," NASA Rep. TND-1208, 1962.
[72] S. F. Schmidt, "State-space techniques applied to the design of a space navigation system," in *Proc. 1962 Joint Automatic Control Conf.*, Paper 11-3.
[73] V. A. Yakubovic, "The solution of certain matrix inequalities in automatic control theory," *Dokl. Akad. Nauk SSSR*, vol. 143, pp. 1304–1307, 1962.
[74] A. E. Bryson and M. Frazier, "Smoothing for linear and nonlinear dynamic systems," Aeronaut. Syst. Div., Wright-Patterson AFB, Ohio, Tech. Rep. ASD-TDR-63-119, Feb. 1963.
[75] R. E. Kalman, "Lyapunov functions for the problem of Lur'e in automatic control," *Proc. Nat. Acad. Sci.*, vol. 49, pp. 201–205, Feb. 1963.
[76] ——, "Mathematical description of linear dynamical systems," *SIAM J. Contr.*, vol. 1, pp. 152–192, 1963.
[77] ——, "On a new characterization of linear passive systems," in *Proc. 1st Annu. Allerton Conf. Circuit and System Theory*, Nov. 1963, pp. 456–470.
[78] A. G. J. MacFarlane, "An eigenvector solution of the optimal linear regulator," *J. Electron. Contr.*, vol. 14, pp. 643–654, June 1963.
[79] E. A. Robinson, "Mathematical development of discrete filters for detection of nuclear explosions," *J. Geophys. Res.*, vol. 68, pp. 5559–5567, 1963.
[80] P. Swerling, "Comment on 'A statistical optimizing navigation procedure for space flight'," *AIAA J.* vol. 1, p. 1968, Aug. 1963.
[81] P. Whittle, "On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix," *Biometrika*, vol. 50, pp.129–134, 1963.
[82] K. Astrom and S. Wensmark, "Numerical identification of stationary time-series," in *Proc. 6th Int. Instrumentation and Measurements Congr.*, Sept. 1964.
[83] R. H. Battin, *Astronautical Guidance.* New York: McGraw-Hill, 1964.
[84] R. B. Blackman, "Methods of orbit refinement," *Bell Syst. Tech. J.*, vol. 43, pp. 885–909, May 1964.
[84a] H. Cox, "On the estimation of state variables and parameters for noisy dynamic systems," *IEEE Trans. Automat. Contr.*, vol. AC-9, pp. 5–12, Jan. 1964.
[85] R. E. Kalman, "When is a linear control system optimal?," *Trans. ASME, Ser. D, J. Basic Eng.*, vol. 86, pp. 51–60, June 1964.
[86] H. J. Kushner, "On differential equations satisfied by conditional probability densities of Markov processes," *SIAM J. Contr.*, vol. 2, pp. 106–119, 1964.
[87] V. M. Popov, "Hyperstability and optimality of automatic systems with several control functions," *Rev. Roum. Sci. Tech.*, vol. 9, pp. 629–690, 1964.
[87a] ——, "Incompletely controllable positive systems and applications to optimization and stability of automatic control systems," *Rev. Roum. Sci. Tech., Electrotech. Energ.*, vol. 12, pp. 337–357, 1967.
[88] G. L. Smith, "Multivariable linear filter theory applied to space vehicle guidance," *SIAM J. Contr.*, vol. 2, pp. 19–32, 1964.
[89] R. L. Stratonovich and Yu. G. Sosulin, "Optimal detection of a Markov process in noise," *Eng. Cybern.*, vol. 6, pp. 7–19, Oct. 1964.
[90] A. E. Bryson and D. E. Johansen, "Linear filtering for time-varying systems using measurements containing colored noise," *IEEE Trans. Automat. Contr.*, vol. AC-10, pp. 4–10, Jan. 1965.
[91] R. S. Bucy, "Nonlinear filtering theory," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-10, p. 198, Apr. 1965.
[92] H. H. Rosenbrock, "On the connection between discrete linear filters and some formulae of Gauss," in *Act. Congr. Automatique Théorique.* Paris: Dunod, 1965.
[93] F. C. Schweppe, "Evaluation of likelihood functions for Gaussian signals," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 61–70, 1965.
[94] E. B. Stear, "Shaping filters for stochastic processes," in *Modern Control Systems Theory*, C. T. Leondes, Ed. New York: McGraw-Hill, 1965, pp. 121–155.
[95] P. Whittle, "Recursive relations for predictors of non-stationary processes," *J. Roy. Statist. Soc., Ser. B*, vol. 27, pp. 525–532, 1965.
[96] R. A. Wiggins and E. A. Robinson, "Recursive solution to the multichannel filtering problem," *J. Geophys. Res.*, vol. 70, pp. 1885–1891, Apr. 1965.
[97] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *SIAM J. Contr.*, vol. 2, pp. 347–369, 1965.
[98] B. D. O. Anderson, "Time-varying spectral factorization," Stanford Electron. Lab., Stanford, Calif., Tech. Rep. SEL-66-107, Oct. 1966.
[98a] P. Businger and G. H. Golub, "Linear least-squares solutions by Householder transformation," *Math. Comput.*, vol. 20, pp. 325–328, 1966.
[99] R. E. Mortensen, "Optimal control of continuous-time stochastic systems," Ph.D. dissertation, Univ. California, Berkeley, 1966.
[100] J. E. Potter, "Matrix quadratic solutions," *SIAM J. Appl. Math.*, vol. 14, pp. 496–501, May 1966.
[101] A. N. Shiryaev, "Stochastic equations of nonlinear filtration for purely discontinuous Markov processes," *Probl. Peredach. Inform.*, vol. 2, pp. 3–22, 1966.

[102] R. L. Stratonovich and Yu. G. Sosulin, "Optimum reception of signals in nonGaussian noise," *Radio Eng. Electron.*, vol. 11, pp. 497–507, Apr. 1966.
[103] D. C. Youla, "The synthesis of linear dynamical systems from prescribed weighting patterns," *SIAM J. Appl. Math.*, vol. 14, pp. 527–549, May 1966.
[104] B. D. O. Anderson, "An algebraic solution to the spectral factorization problem," *IEEE Trans. Automat. Contr.*, vol. AC-12, pp. 410–414, Aug. 1967.
[105] B. D. O. Anderson and J. B. Moore, "Solution of a time-varying Wiener filtering problem," *Electron. Lett.*, vol. 3, pp. 562–563, Dec. 1967.
[106] K. J. Astrom, *Introduction to Stochastic Control Theory*. New York: Academic Press, 1967.
[106a] A. Bjorck and G. H. Golub, "Iterative refinement of linear least-squares solutions by Householder transformation," *BIT*, vol. 7, pp. 322–337, 1967.
[107] J. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Annu. Meet. Soc. Explor. Geophys.*, 1967.
[108] T. E. Duncan, "Probability densities for diffusion processes with applications to nonlinear filtering theory and detection theory," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, Calif., June 1967.
[109] H. J. Kushner, *Stochastic Stability and Control*. New York: Academic, 1967.
[110] E. A. Robinson, *Multichannel Time-Series Analysis with Digital Computer Programs*. San Francisco, Calif., Holden-Day, 1967.
[111] W. M. Wonham, "Lecture notes on stochastic optimal control," Div. Appl. Math., Brown Univ., Providence, R.I., Rep. 67-1, 1967.
[112] C. Bruni, A. Isidori, and A. Ruberti, "A method of factorization of the impulse-response matrix," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-13, pp. 739–741, Dec. 1968.
[113] R. S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*. New York: Wiley, 1968.
[114] L. D. Collins, "Realizable whitening filters and state-variable realizations," *Proc. IEEE* (Lett.), vol. 56, 100–101, Jan. 1968.
[115] T. E. Duncan, "Evaluation of likelihood functions," *Inform. Contr.*, vol. 13, pp. 62–74, July 1968.
[116] T. Kailath, "An innovations approach to least-squares estimation—Part I: Linear filtering in additive white noise," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 646–655, Dec. 1968.
[117] T. Kailath and P. Frost, "An innovations approach to least-squares estimation, Part II: Linear smoothing in additive white noise," *IEEE Trans. Automat. Cont.*, vol. AC-13, pp. 655–660, Dec. 1968.
[118] G. Kallianpur and C. Striebel, "Estimation of stochastic systems: arbitrary system process with additive white observation errors," *Ann. Math. Statist.*, vol. 39, pp. 785–801, 1969.
[119] R. E. Kalman, "Lectures on controllability and observability," Lecture Notes, CIME, Bologna, 1968.
[120] R. S. Lipt'ser and A. N. Shiryaev, "Nonlinear filtering of Markov diffusion processes," *Proc. Steklov Inst.* Math. (English Transl.), vol. 104, pp. 163–218, 1968.
[121] ——, "Nonlinear interpolation of Markov diffusion processes," *Theory Prob. Appl.* (USSR). vol. 13, pp. 564–583, 1968.
[122] J. B. Moore and B. D. O. Anderson, "Extensions of quadratic minimization theory, I," *Int. J. Contr.*, vol. 7, pp. 465–472, 1968.
[123] F. C. Schweppe and H. K. Knudsen, "The theory of amorphous cloud trajectory prediction," *IEEE Trans. Inform Theory*, vol. IT-14, pp. 415–427, May 1968.
[124] E. Stear and A. Stubberud, "Optimal filtering for Gauss–Markov noise," *Int. J. Contr.*, vol. 8, pp. 123–130, 1968.
[125] W. G. Tuel, "Computer algorithm for spectral factorization of rational matrices," *IBM J. Res. Develop.*, vol. 12, pp. 163–170, Mar. 1968.
[126] W. M. Wonham, "On a matrix Riccati equation of stochastic control," *SIAM J. Contr.*, vol. 6, pp. 681–697, Nov. 1968.
[127] B. D. O. Anderson, J. B. Moore, and S. G. Loo, "Spectral factorization of time-varying covariance functions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 550–557, Sept. 1969.
[128] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*. Waltham, Mass.; Blaisdell, 1969.
[128a] P. Dyer and S. McReynolds, "Extension of square-root filtering to include process noise," *J. Optimiz. Theory Appl.*, vol. 3, pp. 444–459, 1969.
[129] R. E. Kalman, P. Falb, and M. A. Arbib, *Topics in Mathematical System Theory*. New York: McGraw-Hill, 1969.
[129a] N. Morrison, *Introduction to Sequential Smoothing and Prediction*. New York: McGraw-Hill, 1969.
[130] N. E. Nahi, "Optimum recursive estimation with uncertain observation," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 457–462, July 1969.
[131] J. Rissanen and L. Barbosa, "Properties of infinite covariance matrices and stability of optimum predictors," *Inform. Sci.*, vol. 1, pp. 221–236, 1969.
[132] W. M. Wonham, "Random differential equations in control theory," in *Probabilistic Methods in Applied Mathematics*, A. T. Bharucha-Reid, Ed. New York: Academic Press, 1969, ch. 2.
[133] L. E. Zachrisson, "On optimal smoothing of continuous-time Kalman processes," *Inform. Sci.*, vol. 1, pp. 143–172, 1969.
[134] L. H. Brandenburg and M. E. Meadows, "Shaping filter representation of nonstationary colored noise," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 26–31, Jan. 1971; see also L. Brandenburg, Ph.D. dissertation, Columbia Univ., New York, June 1970.
[135] R. W. Brockett, *Finite-Dimensional Linear Systems*. New York: Wiley, 1970.
[136] R. S. Bucy, "Linear and nonlinear filtering," *Proc. IEEE*, vol. 58, pp. 854–864, June 1970.
[137] A. Gersho and D. J. Goodman, "Projecting filters for recursive prediction of discrete-time processes," *Bell Syst. Tech. J.*, vol. 49, pp. 2377–2403, Nov. 1970.
[138] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1970.
[139] C. T. Leondes, Ed., "Theory and applications of Kalman filtering," NATO Advanced Group for Aerospace R&D, AGARDOgraph 139, Feb. 1970.
[140] J. B. Moore and B. D. O. Anderson, "Spectral factorization of Time-varying covariance functions," *Math. Syst. Theory*, vol. 4, pp. 10–23, 1970.
[141] R. L. Stratonovich, "Detection and estimation of signals in noise when one or both are non-Gaussian," *Proc. IEEE*, vol. 58, pp. 670–679, May 1970.
[142] H. Akaike, "Autoregressive model fitting for control," *Ann. Inst. Statist. Math.*, vol. 23, pp. 163–180, 1971.
[143] B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*. Englewood Cliffs, N.J., Prentice-Hall, 1971.
[144] ——, "The Kalman–Bucy filter as a true time-varying Wiener filter," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-1, pp. 119–128, Apr. 1971.
[145] M. Athans, Ed., *Special Issue on Linear-Quadratic-Gaussian Problem, IEEE Trans. Automat. Contr.*, vol. AC-16, Dec. 1971.
[146] R. Bellman and E. D. Denman, Eds., "Invariant imbedding," in *Lecture Notes in Operations Research*, vol. 52. New York: Springer, 1971.
[147] G. Epstein, "On finite-memory, recursive filters," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-17, pp. 486–487, July 1970; see also *ibid.*, p. 614, Sept. 1971, p. 753, Nov. 1971.
[148] P. G. Kaminski, "Square-root filtering and smoothing for discrete processes," Ph.D. dissertation, Stanford Univ., Stanford, Calif., Sept. 1971.
[148a] P. G. Kaminski, A. E. Bryson, Jr., and S. F. Schmidt, "Discrete square root filtering: A survey of current techniques," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 727–735, Dec. 1971.
[149] T. Kailath and R. A. Geesey, "An innovations approach to least-squares estimation—Part IV: Recursive estimation given lumped covariance functions," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 720–727, Dec. 1971.
[150] D. G. Lainiotis, "Optimal nonlinear estimation," *Int. J. Contr.*, vol. 14, pp. 1137–1148, 1971.
[150a] ——, "Optimal linear smoothing: Continuous-data case," *Int. J. Contr.*, vol. 17, pp. 921–930, May 1973.
[151] K. Martensson, "On the matrix Riccati equation," *Inform. Sci.*, vol. 3, pp. 17–49, 1971.
[151a] J. K. Omura, "Optimal receiver design for convolutional codes and channels with memory via control theoretical concepts," *Inform. Sci.*, vol. 3, pp. 243–266, 1971.
[152] I. B. Rhodes, "A tutorial introduction to estimation and filtering," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 688–707, Dec. 1971.
[153] A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communication and Control*. New York: McGraw-Hill, 1971. Reviewed by K. Yao, *IEEE Trans. Inform. Theory* (Book Rev.), vol. IT-19, pp. 374–376, May 1973.
[154] M. D. Srinath and P. K. Rajasekaran, "Estimation of randomly occurring stochastic signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-17, p. 206, Mar. 1971.
[155] P. Swerling, "Modern state estimation methods from the viewpoint of the method of least squares," *IEEE Trans. Automat. Contr.* vol. AC-16, pp. 707–720, Dec. 1971.
[156] A. van den Bos, "Alternative interpretation of maximum entropy spectral analysis," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-17, pp. 493–494, July 1971.
[157] H. L. Van Trees, *Detection, Estimation, Modulation Theory, Pt. III—Radar–Sonar Signal Processing and Gaussian Signals in Noise*. New York: Wiley, 1971.
[158] J. C. Willems, "Least squares stationary optimal control and the algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 621–634, Dec. 1971.
[158a] ——, "Dissipative dynamical systems, Part II: Linear systems with quadratic supply rates," *Arch. Ration. Mech. Anal.*, vol. 45, pp. 352–393, 1972.
[159] M. G. Wood, J. B. Moore, and B. D. O. Anderson, "Study of an

integral equation arising in detection theory," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 677–686, Nov. 1971.
[160] A. V. Balakrishnan, "System theory and stochastic optimization," in *Proc. NATO 1972 Advanced Study Institute on Network and Signal Theory*, Bournemouth, England, 1972.
[160a] A. F. Fath, "Computational aspects of the linear optimal regulator problem," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 547–550, Oct. 1969; also, A. E. Bryson and W. E. Hall, "Optimal control and filter synthesis by Eigenvector decomposition," Dep. Aeron. and Astron., Stanford Univ., Stanford, Calif., Rep. 436, Dec. 1971.
[161] J. L. Casti, R. E. Kalaba, and V. K. Murthy, "A new initial-value method for on-line filtering and estimation," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 515–518, July 1972.
[162] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, "Methods for modifying matrix factorizations," Comput. Sci. Dep., Stanford Univ., Stanford, Calif., Rep. CS-72-322, Nov. 1972.
[163] K. L. Hitz and B. D. O. Anderson, "Iterative method of computing the limiting solution of the matrix Riccati differential equation," *Proc. Inst. Elec. Eng.* vol. 119, pp. 1402–1406, Sept. 1972.
[164] F. Itakura, "Extraction of feature parameters of speech by statistical methods," in *Proc. 8th Symp. Speech Information Processing*, Feb. 1972.
[165] T. Kailath, "A note on least-squares estimation by the innovations method," *J. SIAM Contr.*, vol. 10, pp. 477–486, Aug. 1972.
[166] ——, "Some Chandrasekhar-type algorithms for quadratic regulator problems," in *Proc. IEEE Conf. Decision and Control and 11th Symp. Adaptive Processes*, Dec. 1972, pp. 219–223.
[167] V. Kučera, "A contribution to matrix quadratic equations," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 344–346, June 1972.
[167a] ——, "A review of the matrix Riccati equation," *Kybernetika*, vol. 9, pp. 42–61, 1973.
[168] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. New York: Wiley, 1972.
[169] J. T. H. Lo, "Finite-dimensional sensor orbits and optimal nonlinear filtering," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 583–588, Sept. 1972.
[170] D. Rappaport, "Constant directions of the Riccati equation," *Automatica*, vol. 8, pp. 175–186, Mar. 1972. See also *IEEE Trans. Automat. Contr.*, vol. AC-15, pp. 535–540, Oct. 1970.
[171] J. M. Rodriguez-Canabal, "The geometry of the Riccati equation," Ph.D. dissertation, Univ. Southern California, Los Angeles, June 1972.
[172] D. L. Snyder, "Filtering and detection for doubly stochastic Poisson processes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 91–101, Jan. 1972.
[173] H. Wakita, "Estimation of the vocal tract shape by optimal inverse filtering," Speech Commun. Res. Lab., Inc., Santa Barbara, Calif., Mono. 9, July 1972. See also *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, Oct. 1973.
[174] G. T. Wilson, "The factorization of matrical spectral densities," *SIAM J. Appl. Math.*, vol. 23, pp. 420–426, Dec. 1972.
[175] H. Aasnaes and T. Kailath, "An innovations approach to least-squares estimation—Pt VII: Some applications of vector autoregressive-moving average models," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 601–607, Dec. 1973.
[176] H. Aasnaes and T. Kailath, "Robustness of linear-least squares filtering algorithms," in *Proc. 1973 Joint Automatic Control Conf.* See also *IEEE Trans. Automat. Contr.*, vol. AC-19, June 1974.
[177] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis—A Modern Systems Theory Approach*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.
[178] J. A. Edward and M. M. Fitelson, "Notes on maximum-entropy processing, *IEEE Trans. Inform. Theory*. (Corresp.), vol. IT-19, pp. 232–234, Mar. 1973.
[179] M. Gevers and T. Kailath, "Constant, predictable and degenerate directions of the discrete-time Riccati equation," *Automatica*, vol. 9, pp. 699–712, Nov. 1973.
[180] ——, "An innovations approach to least-squares estimation—Part VI: Discrete-time innovations representations and recursive estimation," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 588–600, Dec. 1973.
[181] T. Kailath and R. Geesey, "An innovations approach to least-squares estimation—Part V: Innovations representations and recursive estimation in colored noise," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 435–453, Oct. 1973.
[182] R. H. Jones, "Autoregressive spectrum estimation," in *Proc. Amer. Meteorol. Soc. 3rd Conf. Probability and Statistics in Atmospheric Science*, 1973.
[183] T. Kailath, "Some new algorithms for recursive estimation in constant linear systems," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 750–760, Nov. 1973.
[184] T. Kailath, M. Morf, and G. S. Sidhu, "Some new algorithms for recursive estimation in constant discrete-time linear systems," in *Proc. 7th Princeton Symp. Information and System Science*, 1973. See also *IEEE Trans. Automat. Contr.*, vol. AC-19, Aug. 1974.
[185] G. S. Sidhu, T. Kailath, and M. Morf, "Development of fast algorithms via innovations decompositions," in *Proc. 7th Hawaii Int. Conf. Syst. Sci.*, Honolulu, Hawaii, Jan. 1974.
[186] R. Sh. Lipt'ser and A. N. Shiryaev, "Statistics of conditionally Gaussian random sequences," in *Proc. 6th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. III. Berkeley, Calif.: Univ. California Press, 1973, pp. 389–422.
[187] B. P. Molinari, "The stabilizing solution of the algebraic Riccati equations," *SIAM J. Contr.*, vol. 11, pp. 262–272, May 1973.
[188] ——, "Equivalence relations for the algebraic Riccati equation," *SIAM J. Contr.*, vol. 11, pp. 272–286, May 1973.
[189] J. B. Moore and P. Hetrakul, "Optimal demodulation of PAM signals," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 188–197, Mar. 1973.
[190] H. J. Payne and L. M. Silverman, "On the discrete time algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 226–234, June 1973.
[190a] M. Aoki, "On the subspaces associated with partial reconstruction of state vectors, the structure algorithm and the predictable directions of Riccati equations," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 399–400, Aug. 1973.
[191] M. Morf, "Discrete-time multivariable systems," Ph.D. dissertation, Stanford Univ., Stanford, Calif., 1974.
[192] J. Rissanen, "A fast algorithm for optimum predictors," *IEEE Trans. Automat. Contr.*, vol. AC-18, p. 555, Oct. 1973.
[193] ——, "Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factorizing positive matrix polynomials," *Math. Comput.*, vol. 27, pp. 147–154, 1973.
[193a] F. C. Schweppe, *Uncertain Dynamic Systems*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.
[193b] A. V. Balakrishnan, "Stochastic differential systems, Pt. I," in *Lecture Notes in Economics and Mathematical Systems*, vol. 84. New York: Springer, 1973.
[193c] H. J. Payne and L. Silverman, "Matrix Riccati equations and system structure," in *Proc. IEEE Conf. on Decision and Control*, pp. 558–563, Dec. 1973.
[193d] M. Morf and T. Kailath, "Square-root algorithms for linear least-squares estimation and control," in *Proc. 8th Princeton Symp. Information and System Science*, 1974.

C. Some Early Mathematical Work

[194] G. Galileo, "Dialogo Sopra i due Massimi Sestemi del Mondo: Tolemaico e Copernicano. Florence: Landini, 1632 (Transl.: Berkeley, Calif.: Univ California Press, 1953.)
[195] Jacopo Francesco, Count Riccati, "Animadversationes in aequationes differentiales secundi gradus," in *Actorum Eruditorum quae Lipsiae publicantur*, Suppl. 8, pp. 66–73, 1724.
[196] R. Adrain, "Research concerning the probabilities of the errors which happen in making observations," *Analyst*, vol. 1, pp. 193–209, 1808.
[197] C. F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*, Hamburg, 1809 (Transl.: Dover, New York: 1963).
[198] A. M. Legendre, "Methode des moindres quarrés, pour trouver le milieu le plus probable entre les résultats de differentes observations," *Mem. Inst. France*, pp. 149–154, 1810.
[199] G. H. Hardy, "The mean value of the modulus of an analytic function," *Proc. London Math. Soc.*, vol. 14, pp. 269–277, 1915.
[200] G. Szegö, "Ein Grenszwertsatz uber die Toeplitzschen Determinanten einer reellen positiven Funktion," *Math. Ann.*, vol. 76, pp. 490–503, 1915.
[201] ——, "Beitrage zur Theorie der Toeplitzschen Formen," *Math. Z.*, vol. 6, pp. 167–202, 1920.
[202] R. Frisch, "Correlation and scatter in statistical variables," *Nord. Stat. Tidsk.*, vol. 8, pp. 36–102, 1928.
[203] A. N. Kolmogorov, *Foundations of the Theory of Probability*, New York: Springer, 1933 (Transl.: New York: Chelsea, 1950).
[204] R. E. A. C. Paley and N. Wiener, "Fourier transforms in the complex domain," *Amer. Math. Soc. Colloq. Publ.*, vol. 19, 1934.
[205] M. Frechet, *Recherches théoriques modernes sur la théorie des Probabilities*, vol. 1. Paris: Gauthier-Villars, 1937; 2nd ed., 1950.
[206] H. Wold, *A Study in the Analysis of Stationary Time Series*. Uppsala, Sweden: Almqvist and Wiksell, 1938; 2nd. ed., 1954.
[207] A. N. Kolmogorov, "Sur l'interpolation et extrapolation des suites stationnaires," *C. R. Acad. Sci.*, vol. 208, p. 2043, 1939.
[208] G. Szegö, "Orthogonal polynomials," *Amer. Math. Soc. Colloq. Publ.*, vol. 23, 1939; 2nd ed., 1958; 3rd ed. 1967.
[209] A. N. Kolmogorov, "Stationary sequences in Hilbert space" (in Russian), *Bull. Math. Univ. Moscow*, vol. 2, no. 6, 1941 (A transl. by N. Artin is available in many libraries).
[210] ——, "Interpolation and extrapolation of stationary random sequences," *Bull. Acad. Sci. USSR, Ser Math.*, vol. 5, 1941; Transl.: RAND Corp., Santa Monica, Calif., Memo. RM-3090-PR, Apr. 1962).

[211] V. A. Ambartsumian, "Diffuse reflection of light by a foggy medium," *Dokl. Akad. Sci. SSSR*, vol. 38, pp. 229–322, 1943.
[212] R. S. Phillips, "Servomechanisms," M.I.T. Radiation Lab., Cambridge, Mass.: Rep. 372, May 1943; also in *Theory of Servomechanisms*, H. M. James, N. B. Nichols, and R. S. Phillips, Eds. New York: McGraw-Hill, 1957, ch. 7.
[213] J. L. Doob, "The elementary Gaussian processes," *Ann. Math. Statist.*, vol. 15, pp. 229–282, 1944.
[214] M. G. Krein, "On a generalization of some investigations of G. Szegö, W. M. Smirnov, and A. N. Kolmogorov," *Dokl. Akad. Nauk SSSR*, vol. 46, pp. 91–94, 1945.
[215] ——, "On a problem of extrapolation of A. N. Kolmogorov," *Dokl. Akad. Nauk SSSR*, vol. 46, pp. 306–309, 1945.
[216] W. T. Reid, "A matrix differential equation of the Riccati type," *Amer. J. Math.*, vol. 68, pp. 237–246, 1946.
[216a] ——, *Riccati Differential Equations*. New York: Academic Press, 1972.
[217] S. Chandrasekhar, "On the radiative equilibrium of a stellar atmosphere, Pt XXI," *Astrophys. J.*, vol. 106, pp. 152–216, 1947; Pt XXII, *ibid*, vol. 107, pp. 48–72, 1948.
[218] N. Levinson, "The Wiener rms (root-mean-square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, Jan. 1947; reprinted as appendix in [1].
[219] J. L. Doob, "Time series and harmonic analysis," in *Proc. Berkeley Symp. Mathematics, Statistics, and Probability*. Berkeley, Calif.: Univ. California Press, 1949, pp. 303–343.
[220] S. Chandrasekhar, *Radiative Transfer*. Oxford, England: Oxford Univ. Press, 1950; also New York: Dover, 1960.
[221] O. Hanner, "Deterministic and nondeterministic stationary random processes," *Ark. Mat.*, vol. 1, pp. 161–177, 1950.
[222] K. Karhunen, "Über die Struktur Stationärer Zufälliger Funktionen," *Ark. Mat.*, vol. 1, pp. 141–160, 1950 (Transl.: RAND Corp., Memo. RM-3091-PR, Apr. 1962).
[223] U. Grenander, "On Toeplitz forms and stationary processes," *Ark. Mat.*, vol. 1, pp. 555–571, 1952.
[224] I. M. Gel'fand and B. H. Levitan, "On the determination of a differential equation from its spectral function," *Izv. Akad. Nauk SSSR*, vol. 15, pp. 309–360, 1951 (Transl.: *Amer. Math. Soc. Transl.*, Ser. 2, vol. 1, pp. 253–304, 1955).
[225] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1953.
[226] M. G. Krein, "Some problems of the effective determination of a nonhomogeneous string by means of its spectral function," *Dokl. Akad. Nauk SSSR*, vol. 93, pp. 617–620, 1953.
[227] ——, "On a fundamental approximation problem in the theory of extrapolation and filtration of stationary processes," *Dokl. Akad. Nauk SSSR*, vol. 94, pp. 13–16, 1954 (Transl.: *Select. Transl. Probl. Math. Statist.*, vol. 4, pp. 127–131, 1964).
[228] ——, "On integral equations governing differential equations of second order," *Dokl. Akad. Nauk SSSR*, vol. 97, pp. 21–24, 1954.
[229] ——, "On a new method of solving linear integral equations of the first and second kinds," *Dokl. Akad. Nauk SSSR*, vol. 100, pp. 413–416, 1955.
[230] ——, "The continuous analogues of theorems on polynomials orthogonal on the unit circle," *Dokl. Akad. Nauk SSSR*, vol. 104, pp. 637–640, 1955.
[231] R. E. Bellman, "Functional equations in the theory of dynamic programming, VII: A partial differential equation for the Fredholm resolvent," *Proc. Amer. Math. Soc.*, vol. 8, pp 435–440, 1957.
[232] K. M. Case, "On Wiener–Hopf equations," *Ann. Phys.*, vol. 2, pp. 384–405, 1957.
[233] E. G. Gladyshev, "On multidimensional stationary random processes," *Theory Prob. Appl.* (in Russian), vol. 3, pp. 425–428, 1958.
[234] I. C. Gohberg and M. G. Krein, "Systems of integral equations on a half-axis with kernels depending on the difference of the arguments," *Usp. Mat. Nauk*, vol. 13, pp. 3–72, 1958.
[235] M. G. Krein, "Integral equations on a half-axis with kernel depending on the difference of the arguments," *Usp. Mat. Nauk*, vol. 13, pp. 3–120, 1958.
[236] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. Berkeley, Calif.: Univ. California Press, 1958.
[237] P. Masani, "Cramér's theorem on monotone matrix-valued functions and the Wold decomposition," in *Probability and Statistics*, U. Grenander, Ed. New York: Wiley, 1959, pp. 175–189.
[238] S. Sandor, "Sur l'equation differentielle matricielle de type Riccati," *Bull. Math. Soc. Sci. Math. Phys.*, Revue Physique Roumaine (New Series), vol. 3, pp. 229–249, 1959.
[239] J. Durbin, "The fitting of time-series models," *Rev. Intern. Statist. Inst.*, vol. 28, pp. 233–244, 1960.
[240] G. Baxter, "Polynomials defined by a difference system," *J. Math. Anal. Appl.*, vol. 2, pp. 223–263, 1961.
[241] L. Ya. Geronimus, *Orthogonal Polynomials* (Transl. from the Russian). New York: Consultant's Bureau, 1961.
[242] G. Baxter, "An asymptotic result for the finite predictor," *Math. Scand.*, vol. 10, pp. 137–144, 1962.
[243] J. Hajek, "On linear statistical problems in stochastic processes," *Czech. Math. J.*, vol. 12, pp. 404–444, 1962.
[244] K. Hoffman, *Banach Spaces of Analytic Functions*. Englewood Cliffs, N.J.: Prentice-Hall, 1962.
[245] P. Masani, "Shift-invariant spaces and prediction theory," *Acta Math.*, vol. 107, pp. 275–290, 1962.
[246] P. Masani and J. B. Robertson, "The time-domain analysis of continuous-parameter, weakly stationary stochastic processes," *Pac. J. Math.*, vol. 12, pp. 1361–1378, 1962.
[247] R. M. Redheffer, "On the relation of transmission-line theory to scattering and transfer," *J. Math. Phys.*, vol. 41, pp. 1–41, 1962.
[247a] G. Baxter, "A norm inequality for a 'finite-section' Wiener–Hopf equation," *Ill. J. Math.*, vol. 7, pp. 97–103, 1963.
[248] M. Loève, Probability Theory, 3rd ed. New York: Van Nostrand Reinhold, 1963.
[248a] F. B. Atkinson, *Discrete and Continuous Boundary Problems*. New York: Academic Press, 1964.
[249] A. Devinatz, "Asymptotic estimates for the finite predictor," *Math. Scand.*, vol. 15, pp. 111–120, 1964.
[250] H. Helson, *Lectures on Invariant Subspaces*. New York: Academic Press, 1964.
[251] I. A. Ibragimov, "On the asymptotic behavior of the prediction error," *Theory Prob. Appl. USSR*, vol. 9, pp. 627–633, 1964.
[252] B. Noble, "The numerical solution of nonlinear integral equations and related topics," in *Nonlinear Integral Equations*, P. M. Anselone, Ed. Madison, Wis.: Wisconsin Univ. Press, 1964.
[253] C. M. Deo, "Prediction theory of nonstationary random processes," *Sankhya*, Ser. A, vol. 27, pp. 113–132, 1965.
[253a] J. L. Doob, "Wiener's work in probability theory," *Bull. Amer. Math. Soc.*, vol. 72, no. 1, pt. II, pp. 69–72, Jan. 1966.
[254] H. Kagiwada and R. E. Kalaba, "An initial-value method for Fredholm integral equations of convolution type," RAND Corp. Memo RM-5186-PR, 1966.
[255] P. Masani, "Wiener's contribution to generalized harmonic analysis, prediction theory and filter theory," *Bull. Amer. Math. Soc.*, vol. 72, no. 1, pt. II, pp. 73–125, Jan. 1966.
[256] G. M. Wing, "On certain integral equations reducible to initial value problems," *SIAM Rev.*, vol. 9, pp. 655–670, 1967.
[257] A. Devinatz and M. Shinbrot, "General Wiener–Hopf operators," *Trans. Amer. Math. Soc.*, vol. 145, pp. 467–494, Nov. 1967.
[258] P. L. Duren, *Theory of H^p Spaces*. New York: Academic Press, 1970.
[259] H. Dym and H. P. McKean, "Application of de Branges spaces of integral functions to the prediction of stationary Gaussian processes," *Ill. J. Math.*, vol. 14, pp. 299–343, 1970.
[260] H. Dym and H. P. McKean, "Extrapolation and interpolation of stationary Gaussian processes," *Ann. Math. Statist.*, vol. 41, pp. 1817–1844, Dec. 1970.
[261] B. Sz. Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space*. New York: Academic Press, 1970.
[261a] L. D. Pitt, "On problems of trigonometrical approximation from the theory of stationary Gaussian processes," *J. Multivariable Anal.*, vol. 2, pp. 145–161, June 1972.

D. Canonical Representations, Innovations, Martingales, and All That

[262] V. I. Krylov, "On functions regular in a half-plane," *Mat. Sb.*, vol. 6, pp. 95–138, 1939 (Transl.: *Amer. Math. Soc. Transl.* (2), vol. 32, pp. 37–81, 1963.
[263] H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Princeton, N.J.: Van Nostrand Reinhold, 1945.
[264] A. Beurling, "On two problems concerning linear transformations in Hilbert space," *Acta Math.*, vol. 81, pp. 239–255, 1949.
[265] A. M. Yaglom and M. S. Pinsker, "Random processes with stationary increments of order n," *Dokl. Akad. Nauk SSSR*, vol. 90, pp. 731–733, 1953.
[266] P. Lévy, "Sur une classe de courbes de l'espace de Hilbert et sur une équation intégrale non linéaire," *Ann. Sci. Ec. Norm. Super.*, vol. 73, pp. 121–156, 1956.
[267] N. Wiener and G. Kallianpur, "Nonlinear prediction," Office of Naval Research, Tech. Rep. 1, Cu-2-56-NONR-266, (39)-CIRMIP, Project NR-047-015, 1956.
[268] H. Cramér, "On some classes of nonstationary stochastic processes," in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*. Berkeley, Calif.: Univ. California Press, 1960, pp. 57–78.
[269] T. Hida, "Canonical representations of Gaussian processes and their applications," *Mem. College Sci., Univ. Kyoto, Ser. A*, vol. 33, pp. 109–155, 1960.
[270] L. A. Zadeh, "Time-varying networks, I," *Proc. IRE*, vol. 49, pp. 1488–1502, Oct. 1961.
[271] E. A. Robinson, "Extremal representation of stationary stochastic processes," *Ark. Mat.*, vol. 4, pp. 379–384, 1962.
[272] ——, *Random Wavelets and Cybernetic Wavelets*. New York: Hafner, 1962.
[273] H. Cramér, "Stochastic processes as curves in Hilbert space," *Teor. Veroyat. Primen.*, vol. 9, pp. 169–179, 1964.
[274] P. Lévy, *Processus Stochastiques et Mouvement Brownien*, 2nd ed. Paris: Gauthier-Villars, 1964.

[275] H. Cramér, "A contribution to the multiplicity theory of stochastic processes," in *Proc. 5th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. 2. Berkeley, Calif.: Univ. California Press, 1965, pp. 215–224.
[276] L. A. Shepp, "Radon–Nikodym derivatives of Gaussian measures," *Ann. Math. Statist.*, vol. 37, pp. 321–354, Apr. 1966.
[277] E. Wong and M. Zakai, "On the relation between ordinary and stochastic differential equations and applications to stochastic problems in control theory," in *Proc. 3rd IFAC Congr.* London: Butterworth, 1966.
[277a] E. J. McShane, "Stochastic functional equations: Continuity properties and relation to ordinary equations," in *Control Theory and the Calculus of Variations*, A. V. Balakrishnan, Ed. New York: Academic Press, 1969.
[277b] ——, "Stochastic differential equations and models of random processes," in *Proc. 6th Berkeley Symp. Mathematics, Statistics, and Probability*. Berkeley: Univ. California Press, 1973.
[278] H. Kunita and S. Watanabe, "On square-integrable martingales," *Nagoya Math. J.*, vol. 30, pp. 209–245, Aug. 1967.
[279] P. A. Frost, "Nonlinear estimation in continuous-time systems," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, Calif., June 1968.
[280] I. C. Gohberg and M. G. Krein, "Theory and Applications of Volterra Operators in Hilbert Space," (in Russian). Moscow: Nauka, 1967. (Transl.: *Amer. Math. Soc.*, Providence, R.I., 1970.)
[281] R. Geesey, "Canonical representations of second-order processes with applications," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, Calif., 1968.
[282] M. Hitsuda, "Representation of Gaussian processes equivalent to Wiener processes," *Osaka J. Math.*, vol. 5, pp. 299–312, 1968.
[283] J. M. C. Clark, "Conditions for one-to-one correspondence between an observation process and its innovations," Cent. Comput. Automat., Imperial College, London, Tech. Rep. 1, 1969.
[284] R. Geesey and T. Kailath, "Comments on 'The relationship of alternate state-space representations in linear problems'," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-14, pp. 113–114, Feb. 1969.
[285] R. Geesey and T. Kailath, "Applications of the canonical representation to estimation and detection of colored noise," in *Proc. Symp. Computer Processing in Communications*. Brooklyn, N.Y.: Polytechnic Inst. Brooklyn Press, Apr. 1969.
[286] I. M. Gel'fand and N. Ya. Vilenkin, *Generalized Functions*. New York: Academic Press, 1968.
[287] T. Kailath and R. Geesey, "Covariance factorization—an explication via examples," in *Proc. 2nd Asilomar Conf. Systems Science*, 1968.
[288] A. Schumitzky, "On the equivalence between matrix Riccati equations and Fredholm resolvents," *J. Comp. Syst. Sci.*, vol. 2, pp. 76–87, June 1968.
[288a] A. McNabb and A. Schumitzky, "Factorization of operators, Part III: Initial value methods for linear two-point boundary value problems," *J. Math. Anal. Appl.*, vol. 31, pp. 391–405, Aug. 1970.
[289] E. R. Berlekamp, *Algebraic Coding Theory*. New York: McGraw-Hill, 1969.
[290] T. Kailath, "Application of a resolvent identity to a linear smoothing problem," *SIAM J. Contr.*, vol. 7, pp. 68–74, Feb. 1969.
[291] ——, "A general likelihood-ratio formula for random signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 350–361, May 1969.
[292] ——, "Fredholm resolvents, Wiener–Hopf equations, and Riccati differential equations," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 665–672, Nov. 1969.
[293] T. E. Duncan, "On the absolute continuity of measures," *Ann. Math. Statist.*, vol. 41, pp. 30–38, 1970.
[294] P. A. Frost, "The innovations process and its application to nonlinear estimation and detection of signals in additive white noise," in *Proc. Univ. Missouri, Rolla–M. J. Kelly Communications Conf.*, Oct. 1970, pp. 7.3.1–7.3.6. See also *Proc. 4th Princeton Symp. on Information and System Science*, 1970.
[295] T. Hida, *Stationary Stochastic Processes*. Princeton, N.J.: Princeton Univ. Press, 1970.
[296] T. Kailath, "Likelihood ratios for Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 276–288, May 1970.
[297] ——, "A further note on a general like.ihood formula for random signals in Gaussian noise," *IEEE Trans. on Inform. Theory*, vol. IT-16, pp. 393–396, July 1970.
[298] ——, "The innovations approach to detection and estimation theory," in *Proc. IEEE*, vol. 58, pp. 680–695, May 1970.
[299] B. D. O. Anderson and T. Kailath, "The choice of signal process models in Kalman–Bucy filtering," *J. Math. Anal. Appl.*, vol. 35, pp. 659–668, Sept. 1971.
[300] H. Cramér, *Structural and Statistical Problems for a Class of Stochastic Processes*. Princeton, N.J.: Princeton Univ. Press, 1971.
[301] P. A. Frost and T. Kailath, "An innovations approach to least-squares estimation—Part III: Nonlinear estimation in white Gaussian noise," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 217–226, June 1971.
[302] P. A. Frost, "Estimation and detection for a simple class of conditionally independent-increment processes," in *Proc. IEEE Decision and Control Conf.*, 1971. Also published as lecture notes for Washington University summer course on current trends in automatic control, St. Louis, Mo., 1970.
[303] T. T. Kadota, M. Zakai, and J. Ziv, "Mutual information of the white Gaussian channel with and without feedback," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 368–371, July 1971.
[304] T. Kailath, "Some extensions of the innovations theorem," *Bell Syst. Tech. J.*, vol. 50, pp. 1487–1494, Apr. 1971.
[305] M. M. Rao, "Local functionals and generalized random fields with independent values," *Theory Prob. Appl.* (in Russian), vol. 16, pp. 457–473, 1971.
[306] J. Rissanen, "Recursive identification of linear systems," *SIAM J. Contr.*, vol. 9, pp. 420–430, Aug. 1971.
[307] J. Baras and R. Brockett, "H^2-functions and infinite-dimensional realization theory," in *Proc. IEEE Decision and Control Conf.*, 1972, pp. 355–360; also, Ph.D. dissertation, Harvard Univ., Cambridge, Mass., Sept. 1973.
[308] P. Bremaud, "A Martingale approach to point processes," Ph.D. dissertation, Univ. Calif., Berkeley, Aug. 1972; also Electron. Res. Lab., Rep. M345.
[309] M. Fujisaki, G. Kallianpur, and H. Kunita, "Stochastic differential equations for the nonlinear filtering problem," *Osaka J. Math.*, vol. 9, pp. 19–40, 1972.
[310] T. Kailath, R. Geesey, and H. Weinert, "Some relations between RKHS norms, and Fredholm equations innovation representations," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 341–348, May 1972.
[311] T. Kailath and D. Duttweiler, "An RKHS approach to detection and estimation problems—Part III: Generalized innovations representations and a likelihood-ratio formula," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 730–745, Nov. 1972.
[312] H. Dym and H. McKean, *Fourier Series and Integrals*. New York: Academic Press, 1972.
[313] J. Rissanen and T. Kailath, "Partial realization of stochastic processes," *Automatica*, vol. 8, pp. 380–386, July 1972.
[314] Yu. A. Rozanov, "On nonanticipative linear transformations of Gaussian processes with equivalent distributions," *Nagoya Math. J.*, vol. 47, pp. 227–235, 1972.
[315] E. Wong, *Stochastic Processes in Information and Communication Systems*. New York: McGraw-Hill, 1971. Reviewed by F. J. Beutler, *IEEE Trans. Inform. Theory* (Book Rev.), vol. IT-18, pp. 827–828, Nov. 1972.
[316] H. Akaike, "Markovian representation of stochastic processes by canonical variables," *SIAM J. Contr.*, to be published, 1973.
[316a] ——, "Stochastic theory of minimal realizations," *IEEE Trans. Automat. Contr.*, vol. AC-19, to be published.
[317] B. D. O. Anderson, "Algebraic properties of minimal degree spectral factors," *Automatica*, vol. 9, pp. 491–500, 1973.
[318] A. Ephremides and L. Brandenburg, "On the reconstruction error of sampled data estimates," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-19, pp. 365–367, May 1973.
[319] A. Ephremides and J. B. Thomas, *Random Processes—Multiplicity Theory and Canonical Decompositions*. Stroudsburg, Pa.: Dowden, Hutchinson and Ross, 1973.
[320] M. Hitsuda, "Multiplicity of some classes of Gaussian processes," *Nagoya J. Math.*, to be published, 1974.
[321] G. Kallianpur and H. Oodaira, "Nonanticipative representations of equivalent Gaussian processes," *Ann. Prob.*, vol. 1, pp. 104–122, 1973.
[322] T. Kailath and A. Segall, "A further note on innovations, martingales and nonlinear estimation," in *Proc. IEEE Decision and Control Conf.*, 1973.
[323] H. P. McKean, "Geometry of differential space," *Ann. Prob.*, vol. 1, pp. 197–206, Apr. 1973.
[324] P. A. Meyer, "Sur un problème de filtration," Séminaire de probabilités, Pt VII, *Lecture Notes in Mathematics*, vol. 321. New York: Springer, 1973, pp. 223–247.
[325] Yu. A. Rozanov, "Innovations and nonanticipative processes," in *Multivariate Analysis*, vol. 3, P. R. Krishnaiah, Ed. New York: Academic Press, 1973.
[326] A. Segall, "A Martingale approach to modeling, estimation and detection of jump processes," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, Calif., Aug. 1973.
[327] E. Wong, "Recent progress in stochastic processes—a survey," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 262–275, May 1973.
[327a] J. H. Van Schuppen, "Estimation theory for continuous-time processes, a martingale approach," Ph.D. dissertation, Dep. Elec. Eng., Univ. California, Berkeley, Sept. 1973.
[327b] R. Boel, P. Varaiya, and E. Wong, "Martingales on jump processes, I: Representation results, II: Applications," Univ. of California, Berkeley, ERL Memos 407 and 409, Oct. 1973; also, submitted to *SIAM J. Contr.*

[327c] L. H. Brandenburg, "Covariance factorization: Some unified results encompassing both stationary and nonstationary processes," to appear in *IEEE Trans. Inform. Theory.*

E. Miscellaneous

1) *Series Expansion (further references can be found in [327])*

[328] D. D. Kosambi, "Statistics in function space," *J. Indian Math. Soc.* vol. 7, pp. 76–88, 1943.
[329] M. Loève, "Sur les functions aléatories stationnaires de second order," *Rev. Sci.*, vol. 83, pp. 297–310, 1945; see also *Comptes Rénd.*, vol. 220, p. 380, 1945, and vol. 222, p. 489, 1946.
[330] M. Kac and A. J. F. Siegert, "On the theory of noise in radio receivers with square-law detectors," *J. App. Phys.*, vol. 18, pp. 383–397, Apr. 1947; see also, *Ann. Math. Statist.*, vol. 18, pp. 438–442, 1947.
[331] K. Karhunen, "Über Lineare Methoden in der Wahrscheinlichkeitsrechnung," *Amer. Acad. Sci.*, Fennicade, Ser. A, I, vol. 37, pp. 3-79, 1947; (Transl.: RAND Corp., Santa Monica, Calif., Rep. T-131, Aug. 1960).
[332] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 63, pp. 337–404, May 1950.
[333] U. Grenander, "Stochastic processes and statistical inference," *Ark. Mat.*, vol. 1, pp. 195–277, Oct. 1950.
[334] A. Ya. Povzner, "A class of Hilbert function spaces," *Dokl. Akad. Nauk SSSR*, vol. 68, pp. 817–820, 1949; see also, *ibid.*, vol. 74, pp. 13–17, 1950.
[335] R. C. Davis, "On the theory of prediction of nonstationary stochastic processes," *J. Appl. Phys.*, vol. 23, pp. 1047–1053, Sept. 1952.
[336] D. Slepian, "Estimation of signal parameters in the presence of noise," *IRE Trans. Inform. Theory*, vol. P6IT-3, pp. 68–89, Mar. 1954.
[337] D. C. Youla, "The use of the method of maximum likelihood in estimating continuous-modulated intelligence which has been corrupted by noise," *IRE Trans. Inform. Theory*, vol. P6IT-3, pp. 90–105, Mar. 1954.
[338] I. M. Gel'fand and A. M. Yaglom, "Calculation of the amount of information about a random function contained in another such function," *Usp. Mat. Nauk*, vol. 12, pp. 3–52, 1956.
[339] V. S. Pugachev, "Application of canonic expansions of random functions in determining an optimum linear system," *Automat. Remote Contr.*, vol. 17, pp. 489–499, 1956.
[340] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1958.
[341] W. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise.* New York: McGraw-Hill, 1958.
[342] C. W. Helstrom, *Statistical Theory of Signal Detection.* London: Pergamon Press, 1960; 2nd ed., 1968.
[343] V. N. Tutubalin and M. I. Freidlin, "On the structure of the infinitesimal σ-algebra of a Gaussian process," *Teor. Veroyat. Primen.*, vol. 7, pp. 196–199, 1962.
[344] H. P. McKean, "Brownian motion with a several dimensional time," *Teor. Veroyat. Primen.*, vol. 8, pp. 357–378, 1963.
[345] N. Levinson and H. P. McKean, "Weighted trigonometrical approximation on R^1 with application to the germ field of a stationary Gaussian noise," *Acta Math.*, vol. 112, pp. 99–143, 1964.
[346] A. M. Yaglom, "Outline of some topics in linear extrapolation of stationary random processes," in *Proc. 5th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. II. Berkeley, Calif.: Univ. California Press, 1970, pp. 259–278.
[346a] ——, "Strong limit theorems for stochastic processes and orthogonality conditions for probability," in *Proc. Bernoulli, Bayes, Laplace Symp.* Berlin: Springer, 1965, pp. 253–262.
[347] E. Parzen, *Time Series Analysis Papers.* San Francisco: Holden-Day, 1967.
[348] T. T. Kadota, "Optimum estimation of nonstationary Gaussian signals in noise," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 253–257, Mar. 1969.
[349] T. Kailath, "RKHS approach to detection and estimation problems—Part I: Deterministic signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 530–549, Sept. 1971.
[350] R. D. LePage, "Note relating Bochner integrals and reproducing kernels to series expansions on a Gaussian Banach space," *Proc. Amer. Math. Soc.*, vol. 32, pp. 285–289, Mar. 1972. N. Jain and G. Kallianpur, *Proc. Amer. Math. Soc.*, vol. 25, pp. 890–895, 1970.
[350a] E. Lyttkens, "Regression aspects of canonical correlation," *J. Multivariate Anal.*, vol. 2, pp. 418–439, 1972.
[351] S. Cambanis, "A general approach to linear mean-square estimation problems," *IEEE Trans. Inform. Theory*, (Corresp.), vol. IT-19, pp. 110–114, Jan. 1973.
[352] W. A. Gardner, on "A general approach to linear mean-square estimation problems," *IEEE Trans. Inform. Theory*, (Corresp.), vol. IT-19, pp. 114–115, Jan. 1973. See also this issue, pp. 271–274.
[353] T. E. Fortmann and B. D. O. Anderson, "On the approximation of optimal realizable linear filters using a Karhunen-Loeve expansion," *IEEE Trans. Inform. Theory*, (Corresp.), vol. IT-19, pp. 561–564, July 1973.
[353a] V. Belevitch, "On network analysis by polynomial matrices," in *Recent Developments in Network Theory*, S. R. Deards, Ed. Oxford: Pergamon, 1963, pp. 19–30.
[353b] ——, *Classical Network Theory.* San Francisco: Holden-Day, 1968.
[353c] C. Gueguen, A. Fossard, and M. Gauvrit, "Une représentation intermédiaire des systèmes multi-dimensionnels," in *Proc. 1st IFAC Symp. on Multivariable Control*, Dusseldorf, Germany, Oct. 1968.

2) *Linear System Structure*

[354] J. L. Massey, "Shift-register synthesis and BCH decoding," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 122–127, Jan. 1969.
[355] V. M. Popov, "Some properties of control systems with matrix transfer functions," in *Lecture Notes in Mathematics*, vol. 144. Berlin: Springer 1970, pp. 250–261.
[356] H. H. Rosenbrock, *State Space and Multivariable Theory.* New York: Wiley, 1970.
[357] L. Silverman, "Inversion of multivariable linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 270–276, June 1969.
[357a] L. Silverman and H. J. Payne, "Input–output structure of linear systems," *SIAM J. Contr.*, vol. 9, pp. 199–233, May 1971.
[358] S. H. Wang, "Design of linear multivariable systems," Ph.D. dissertation, Univ. Calif., Berkeley, Dec. 1971; also Electron. Res. Lab. Memo. ERL-M309, Oct. 1971.
[358a] A. S. Morse and M. A. Wonham, "Status of noninteracting control," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 568–581, Dec. 1971.
[359] V. M. Popov, "Invariant description of linear, time-invariant controllable system," *SIAM J. Contr.*, vol. 10, pp. 252–264, 1972.
[360] B. Dickinson, M. Morf, and T. Kailath, "A minimal realization algorithm for matrix sequences," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 31–38, Feb. 1974.
[361] G. D. Forney, Jr., "Convolutional codes I: Algebraic structure," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 720–738, Nov. 1970.
[361a] ——, "Structural analysis of convolutional codes via dual codes," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 512–518, July 1973.
[361b] ——, "Minimal bases of rational vector spaces with applications to multivariable linear systems," *SIAM J. Contr.*, 1974.
[362] S. H. Wang and E. J. Davison, "A minimization algorithm for the design of linear multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 220–223, June 1973.
[363] W. A. Wolovich, "The determination of state-space representations for linear multivariable systems," *Automatica*, vol. 9, pp. 97–106, 1973.

3) *Cepstral Analysis*

[364] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes," in *Proc. Symp. Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, ch. 15, pp. 209–243.
[365] B. P. Bogert and J. F. Ossanna, "The heuristics of cepstrum analysis of a stationary complex echoed Gaussian signal in stationary Gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-2, pp. 373–380, July 1966.
[366] R. C. Kemerait and D. G. Childers, "Signal detection and extraction by cepstrum techniques," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 745–759, Nov. 1972.
[367] L. R. Rabiner and C. M. Rader, Eds., *Digital Signal Processing.* New York: IEEE Press, 1972.
[368] T. J. Cohen, "Source-depth determinations using spectral, pseudoautocorrelation and cepstral analysis," *Geophys. J. Roy. Astron. Soc.*, vol. 20, pp. 223–231, 1970.

4) *Historical Surveys*

[369] R. L. Plackett, "A historical note on the method of least squares," *Biometrika*, vol. 36, pp. 458–460, 1949.
[370] O. Neugebauer, *The Exact Sciences in Antiquity.* Princeton, N.J.: Princeton Univ. Press, 1952.
[371] N. Wiener, *I Am a Mathematician.* Cambridge, Mass.: M.I.T. Press, 1956.
[372] C. Eisenhart, "The meaning of 'least' in least squares," *J. Wash. Acad. Sci.*, vol. 54, pp. 24–33, 1964.
[373] H. L. Seal, "The historical development of the Gauss linear model," *Biometrika*, vol. 54, pp. 1–23, 1967.
[374] H. W. Sorenson, "Least-squares estimation: from Gauss to Kalman," *IEEE Spectrum*, vol. 7, pp. 63–68, July 1970.
[375] H. L. Harter, "The method of least squares and some alternatives," Aerospace Res. Lab., Air Force Systems Command, Wright-Patterson AFB, Ohio, Rep. ARL 72-0129, Sept. 1972.
[376] H. W. Sorenson, "Estimation theory: a historical perspective," in *Proc. SW IEEE Conf.*, 1972.

5) *Information-Theoretic Analyses*

[377] I. Vajda, "A contribution to the informational analysis of patterns," in *Methodologies of Pattern Recognition*, M. S. Watanabe, Ed. New York: Academic Press, 1969, pp. 509–519.
[378] S. Arimoto, "Information-theoretical considerations on estimation problems," *Inform. Contr.*, vol. 19, pp. 181–194, 1971.
[379] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 386–391, May 1969.
[380] L. P. Seidman, "Performance limitations and error calculations for parameter estimation," *Proc. IEEE*, vol. 58, pp. 644–652, May 1970.
[381] J. Seidler, "Bounds on the mean-square error and the quality of domain decisions based on mutual information," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 655–665, Nov. 1971.
[382] R. E. Blahut, "An hypothesis-testing approach to information theory," Ph.D. dissertation, Cornell Univ., Ithacar, N.Y., Aug. 1972; Abstract in *IEEE Trans. Inform. Theory* (Dissertation Abstr.), vol. IT-19, p. 253, Mar. 1973.
[383] M. Zakai and J. Ziv, "Lower and upper bounds on the optimal filtering of certain diffusion processes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 325–331, May 1972.
[384] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 275–283, May 1973.

6) *Others*

[385] F. Riesz and B. Sz.-Nagy, *Functional Analysis*. New York: Ungar, 1955.
[386] F. Smithies, *Integral Equations*. New York: Cambridge Univ. Press, 1962.
[387] J. H. Wilkinson and C. Reinsch, *Linear Algebra, Handbook of Automatic Computation*, vol. 2. Berlin: Springer, 1971.
[388] J. R. Klauder and E. G. G. Sudarshan, *Quantum Optics*. New York: Benjamin, 1969.
[389] E. Hopf, "Statistical hydro-dynamics and functional calculus," *J. Ration. Mech. Anal.*, vol. 2, pp. 587–591, 1953.
[390] R. H. Kraichnan, "The closure problem of turbulence theory," in *Proc. Symp. Applied Mathematics*, vol. 13, 1962, pp. 199–225.

A New Estimator for an Unknown Signal Imbedded in Additive Gaussian Noise

MANOUCHEHR MOHAJERI, MEMBER, IEEE

Abstract—**Estimation of an unknown signal observed in the presence of an additive Gaussian noise process is reduced to the problem of estimating an unknown complex parameter. A new class of estimators for an unknown complex parameter is introduced, and their biases and mean-square errors are studied. The performance of a particular member of this class (*c-a* estimator) is compared with that of the maximum-likelihood (ML) estimator, and it is shown that the *c-a* estimator reduces considerably the mean-square error for small values of SNR, at the expense of introducing a small bias. The *c-a* and ML estimators of a complex parameter are applied to the problem of signal estimation, and some interesting numerical results are presented.**

I. Introduction

In many communication problems, such as the discrimination of small-magnitude seismic events, the background noise causes serious difficulties. Seismic discriminants such as complexity (ratio of the signal energy in two different time intervals) and spectral ratio (ratio of the signal energy in two different frequency bands) are powerful tools for discrimination of large-magnitude seismic events [1], [2]. When the event magnitude diminishes, the noise becomes so critical that these discriminants lose their identification capabilities, and therefore one has to search for different means of noise reduction.

Manuscript received April 4, 1972; revised September 6, 1973.
The author was with the Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, Mass. He is now with the Faculty of the Department of Electrical Engineering, Pahlavi University, Shiraz, Iran.

Although various array processing methods have been employed, and significant improvements in the signal-to-noise ratio have been obtained, there still remains a residual noise that needs further reduction [3]. Since by their very nature seismic signals are unknown, this noise reduction should be treated as an *unknown signal* estimation problem.

One widely used technique for estimating an unknown signal is the maximum-likelihood (ML) estimation procedure [4]. When the additive noise process is Gaussian, this estimator has a simple structure, and it chooses the observed waveform as the signal estimate. The ML estimate of an unknown signal, observed in the presence of an additive Gaussian noise process, is an unbiased efficient estimate of the signal, and makes no use of the knowledge of the noise spectrum.

In this paper we introduce a new class of signal estimators which take advantage of the noise spectrum and which are generally biased. The reason for introducing such biased estimators is to reduce the mean-square error at small signal-to-noise ratios. Analysis of the performance of a particular member of this class of estimators shows that, at the expense of introducing a small bias, a considerable reduction in the interval mean-square error, relative to the mean-square error of the ML estimator, is attainable. In seismic signal estimation, the noise process is short-term stationary and this method of estimation proves to be extremely useful. For such a noise process an updated estimate of the noise spectrum is used in the structure of the signal estimator.