An Innovations Approach to Least-Squares Estimation Part I: Linear Filtering in Additive White Noise

THOMAS KAILATH, MEMBER, IEEE

Abstract—**The innovations approach to linear least-squares approximation problems is first to "whiten" the observed data by a causal and invertible operation, and then to treat the resulting simpler white-noise observations problem. This technique was successfully used by Bode and Shannon to obtain a simple derivation of the classical Wiener filtering problem for stationary processes over a semi-infinite interval. Here we shall extend the technique to handle nonstationary continuous-time processes over finite intervals. In Part I we shall apply this method to obtain a simple derivation of the Kalman-Bucy recursive filtering formulas (for both continuous-time and discrete-time processes) and also some minor generalizations thereof.**

I. Introduction

IN THE EARLY 1940's, Kolmogorov [1] and Wiener [2] first discussed problems of linear least-squares estimation for stochastic processes, but by entirely different methods. Kolmogorov [1] studied only discrete-time problems and he solved them by using a simple representation of such processes that was suggested in a 1938 doctoral dissertation by Wold [3]. This representation, which is obtained by a recursive orthonormalization procedure, is known as the Wold decomposition. [The original papers of Kolmogorov and Wold are quite readable, but a more accessible and very readable reference is the monograph by Whittle [4] (especially sec. 3.7).]

On the other hand, Wiener [2] took an almost completely nonprobabilistic approach. He mainly studied continuous-time problems and reduced them to the problem of solving a certain integral equation, the so-called Wiener–Hopf equation, that Wiener and Hopf had solved in 1931 [5] by using some of Wiener's results on harmonic analysis. Though Wiener undertook this work in response to an engineering problem (the design of antiaircraft fire-control systems), his solution was beyond the reach of his engineering colleagues, and his yellow-bound report soon came to be labeled the "Yellow Peril."

In 1950, Bode and Shannon [6] published a different derivation of Wiener's results that was, quite successfully, intended to make them more accessible to engineers. This paper was based on ideas in a classified 1944 report by Blackman, Bode, and Shannon [7]. The same approach was independently discovered by Zadeh (cf. footnote 3 in Zadeh and Ragazzini [8]). However, it is somewhat ironic that these more engineering approaches were found later to be just the continuous-time versions of the original Wold–Kolmogorov technique, which had been developed in a purely mathematical context.

The results in [1]–[7] were all obtained for stationary processes with infinite or semi-infinite observation intervals. The paper of Zadeh and Ragazzini [8] was the first significant attempt to extend the theory. Over the last two decades, various extensions and generalizations have been obtained and many of these have been documented in textbooks, as for example, those of Doob [9], Laning and Battin [10], Pugachev [11], Lee [12], Yaglom [13], Whittle [3], Deutsch [14], Liebelt [15], Balakrishnan [16], Bryson and Ho [17], and others.

In recent years, applications in orbital mechanics and spacecraft tracking have spurred interest in recursive estimation for nonstationary processes over finite-time intervals. Such algorithms were used by Gauss in his numerical calculations of the orbit of the asteroid Ceres, but the modern interest in them is due to Swerling [18] and especially Kalman [19], [20], and Bucy [21], [22]. The great interest in recursive algorithms because of their obvious computational advantages has stimulated a great number of papers on them, providing alternate forms and derivations showing their relationship to more classical parameter estimation techniques (see, for example, the discussions and references in Deutsch [14] and Liebelt [15]). Nevertheless, it seems to us that the original derivations of Kalman [19] and Kalman and Bucy [22] still provide the most insight.

In order to obtain recursive solutions, Kalman and Bucy had to confine themselves to a special class of processes, viz., those that could be generated by passing white noise through a (possibly time-variant) "lumped" linear dynamical system, i.e., a system composed of a finite number of (possibly time-variant) R, L, C elements. (Such processes are sometimes called projections of wide-sense Markov processes, but we shall in the rest of this paper call them "lumped" processes.) They also assumed complete knowledge of this sytsem, thus sidestepping the difficult problem of spectral factorization that had been a stumbling block to the extension of Wiener's classic solution (for semi-infinite observations on a stationary process) to more general situations. In his first paper in 1959, Kalman [19] treated discrete-time processes and obtained a recursive solution by a technique that was essentially the same as Kolmogorov's. In a later paper [20], he extended these results

Manuscript received January 31, 1968. This work was supported by the Applied Mathematics Division of the Air Force Office of Scientific Research under Contract AF 49(638)1517, and by the Joint Services Electronics Program at Stanford University, Stanford, Calif., under Contract Nonr 225(83).

The author is with Stanford University, Stanford, Calif.

to the continuous-time case by the use of a particular limiting technique. This technique, though useful, is somewhat tedious to carry out rigorously. A careful discussion has been given by Wonham [23]. In [22], Kalman and Bucy attacked the continuous-time problem directly. However, they did not use the Wold-Kolmogorov approach because the direct continuous-time analog of Kalman's discrete-time procedure in [19] was hard to see. They therefore returned to the Wiener–Hopf integral equation and showed that (under certain assumptions on the signal and noise processes) the solution to this equation could be expressed in terms of the solution to a nonlinear Riccati differential equation. It is also worth noting that Siegert [24] had carried out essentially the same steps in a different (but mathematically isomorphic) problem.

The chief purpose of Part I is to give a derivation of the Kalman–Bucy results by the Wold–Kolmogorov method, which, for reasons that will be clear later, we shall call the innovations method. Not only does this close a gap in the preceding circle of ideas, but the insight it provides into the proof has also suggested some new results. These include some slight generalizations in the types of processes for which recursive estimation formulas can be obtained, and a very simple and general solution of the so-called smoothing (or interpolation or noncausal filtering) problem. The smoothing problem is one that has been somewhat difficult to solve by the original techniques of Kalman and Bucy, and the solutions that have been obtained are in a somewhat complicated form (see the discussions in Part II [25][1]). Our technique also enables a completely parallel method of attack for the discrete- and continuous-time problems. A new approach to linear estimation with additive colored (nonwhite) noise also follows from the present ideas (Geesey and Kailath [26]).

More strikingly, the innovations technique can also be extended to a large class of nonlinear least-squares problems, viz., those where the observation process is the sum of a non-Gaussian process and additive white Gaussian noise (cf. Kailath and Frost [27] and Frost [28]). The ideas of the present paper have also yielded some general results on the detection of general non-Gaussian signals in additive Gaussian noise (Kailath [29], discrimination between two general Gaussian processes (Kailath and Geesey [30]), and also in certain modeling problems (Kailath and Geesey [31]).

Finally, we should say a word about the level of rigor in the present work. It is difficult to work directly with white noise in a completely satisfactory and rigorous manner—one has usually, especially in the nonlinear case, to work with the integrated white noise. However, in our opinion, the key ideas can always be presented, quite simply, in the white-noise formulation. Then, after some familiarity with the appropriate mathematics has been gained, one can translate the white-noise formulation into the more rigorous (stochastic differential) framework. We shall do this in later papers. The more informal presentation here will, we hope, bring the basic ideas to a wider audience.

[1] This issue, page 655.

II. The Innovations Approach to Linear Least-Squares Estimation

The innovations approach is first to convert the observed process to a white-noise process, to be called the innovations process, by means of a *causal and causally invertible* linear transformation. The point is that the estimation problem is very easy to solve with white-noise observations. The solution to this simplified problem can then be reexpressed in terms of the original observations by means of the inverse of the original "whitening" filter.

This program, used by Bode and Shannon [6] for the stationary process problem with semi-infinite observation time, will now be carried out when the observations are made over a finite-time interval on a continuous-time (possibly nonstationary) stochastic process. Several initial sets of assumptions and several corresponding classes of problems, of varying degrees of generality, can be formulated. For simplicity, however, we shall deal largely with the following additive white-noise problem.

The given observation is a record of the form

$$y(t) = z(t) + v(t), \qquad t \in [a, b) \tag{1}$$

where

$v(\cdot)$ = a sample function of zero-mean white noise with covariance function[2]

$$\overline{v(t)v'(s)} = R(t)\delta(t - s), \qquad R(t) > 0,$$

$z(\cdot)$ = a sample function of a zero-mean "signal" process that has finite variance

$$\operatorname{tr}[\overline{z(t)z'(t)}] < \infty, \qquad t \in [a, b)$$

$[a, b]$ = a finite interval[3] on the real line.

We also assume that the "future" noise $v(\cdot)$ is uncorrelated from the "past" signal $z(\cdot)$, i.e.,

$$\overline{v(t)z'(s)} = 0, \qquad a \leq s < t < b. \tag{2}$$

We shall be interested in the linear least-squares estimate of a related process $x(t)$. Let

$\hat{x}(t \mid b)$ = a linear function of all the data $\{y(s), a \leq s < b\}$ that minimizes the mean-square error $\operatorname{tr}[z(t) - \hat{z}(t \mid b)][z(t) - \hat{z}(t \mid b)]'$. (3)

The corresponding instantaneous estimation error will be written

$$\tilde{z}(t \mid b) = z(t) - \hat{z}(t \mid b), \tilde{z}(t \mid t) = z(t) - \hat{z}(t \mid t). \tag{4}$$

[2] Bars will be used to denote expectations.

[3] The case of an infinite interval requires certain additional assumptions on the signal process such as stationarity, observability of models generating it, etc. Some more specific comments on this point will be made later [after (35)].

When $b=t$, the estimate is usually called the *filtered* estimate, when $b>t$ it is usually called the *smoothed* estimate, and when $b<t$ it is called the *predicted* estimate.

The major tools for the calculation of these estimates will be the following two theorems.

Theorem 1—The Projection Theorem: The best estimate $\hat{z}(t|b)$ is unique and satisfies the conditions

$$\tilde{z}(t|b) \triangleq z(t) - \hat{z}(t|b) \perp y(s), \qquad a \le s < b \quad (5)$$

where

$$u \perp v \quad \text{means that} \quad \overline{uv'} = 0. \quad (6)$$

In words, the instantaneous error is uncorrelated with the observations.

Proof: This theorem, which was used by Kolmogorov [1], is by now fairly well known to engineers and is used in several of the textbooks cited earlier. A brief discussion of the relevant geometric picture is given in Appendix I.

Theorem 2—The Innovations Theorem: The process $\nu(\cdot)$ defined by

$$\nu(t) = y(t) - \hat{z}(t|t) = z(t|t) + v(t), \qquad a \le t < b, \quad (7)$$

and to be called the "innovation process" of $y(\cdot)$, is a *white-noise* process with the *same covariance* as $v(\cdot)$, i.e.,

$$\overline{\nu(t)\nu'(s)} = \overline{v(t)v'(s)}, \qquad a \le t, \quad s < b. \quad (8)$$

Furthermore, $y(\cdot)$ and $\nu(\cdot)$ can be obtained from the other by causal (nonanticipative) linear operations. Therefore, $y(\cdot)$ and $\nu(\cdot)$ are "equivalent" (i.e., they contain the same statistical information) as far as linear operations are concerned.

Proof: The proof will be deferred to Appendix II; however, a few remarks on the theorem and its significance are appropriate here.

Remark 1: The quantity $\nu(t)=y(t)-\hat{z}(t|t)=y(t)-\hat{y}(t|t-)$ may be regarded as defining the "new information" brought by the current observation $y(t)$, being given all the past observations $y(t)$, and the old information deduced therefrom. Therefore, the name "innovation process of $y(\cdot)$" came into being. This term was first used for such processes by Wiener and has since gained wide currency. (A significant generalization, due to Frost [28] and to Kailath [29], of this theorem is that when the white noise $v(\cdot)$ is also Gaussian, but the signal $z(\cdot)$ is non-Gaussian, the innovation process $\nu(\cdot)$ is not only white with the same covariance as $v(\cdot)$, but it is also Gaussian. Applications of this surprising result are given in [27]–[29].)

Remark 2: The fact that $\nu(\cdot)$ is white has been noted before in the special case of lumped signal processes. In this case, the result was probably first noticed by several people (cf. [23], [32]–[34], and unpublished notes of the author and others). However, all their arguments rely, to varying degrees, on the explicit Kalman–Bucy formulas for $\hat{z}(t|t)$. Here we first obtain the result more generally [and with less computation, since we rely only on the projection properties of $\hat{z}(t|t)$], and then use it to obtain the Kalman–Bucy formulas. We note also that the equivalence of $y(\cdot)$ and $\nu(\cdot)$ does not seem to have been explicitly pointed out before, even though, assuming knowledge of the Kalman–Bucy formulas, a proof is immediate (cf. Appendix II-D).

Remark 3: One reason the fact that $\nu(\cdot)$ is white (for lumped processes) may have been known for a long time is that in the discrete-time solution of Kalman [19], $\nu(\cdot)$ is shown to be white (in discrete time) with, however, a different covariance from that of the original noise. The exact formula will be given later (39).

III. Some Applications

We turn now to some applications of these two theorems. First, we present a new derivation of the Kalman–Bucy formulas for filtering of lumped signal processes in white noise. This derivation shows clearly the step at which restriction to such processes is essential to get a recursive solution, and this insight easily yields several (slight) generalizations of the Kalman–Bucy results, including some recent ones due to Kwakernaak [35], Falb [36], Balakrishnan and Lions [37] and Chang [38]. The same techniques apply to discrete-time problems as well. Our new method of proof yields, very simply, a general formula for the smoothed estimate (Part II) [26] and also, more importantly, can be generalized to the nonlinear case (Part III) [27].

A. The Kalman–Bucy Formulas for Recursive Filtering and Prediction

The Kalman–Bucy results are for lumped processes; however, we shall not begin with this assumption, but shall try to see how far we can go without any special assumptions.

We are given $\{y(s)=z(s)+v(s),\ a\le s<t\}$ and wish to calculate the linear least-squares estimate $\hat{x}(t|t)$ of a related random variable $x(t)$.

The first step is to obtain the innovations, which, by Theorem 2, are given by

$$\nu(t) = y(t) - \hat{z}(t/t), \qquad \overline{\nu(t)\nu(s)} = \boldsymbol{R}(t)\delta(t-s). \quad (9)$$

Because the innovations $\nu(\cdot)$ are equivalent to the original observations $y(\cdot)$, we can express $\hat{x}(t|t)$ as

$$\hat{x}(t|t) = \int_a^t \boldsymbol{g}(t, s)\nu(s)ds \quad (10)$$

where the linear filter $\boldsymbol{g}(t, \cdot)$ is to be chosen so that [again using the equivalence of $y(\cdot)$ and $\nu(\cdot)$]

$$x(t) - \hat{x}(t|t) \perp \nu(s), \qquad a \le s \le t. \quad (11)$$

Putting together (9)–(11), we obtain

$$\overline{x(t)\nu'(s)} = \int_a^t \boldsymbol{g}(t, \sigma)\overline{\nu(\sigma)\nu'(s)}d\sigma \quad (12)$$

$$= \boldsymbol{g}(t, s)\boldsymbol{R}(s), \qquad a \le s \le t. \quad (13)$$

It is the last step that justifies the use of the innovation

process $\mathbf{v}(\cdot)$. In (10)–(12), we could equally well have used the original observations $\mathbf{y}(\cdot)$, but now (12), instead of being trivial, becomes the Wiener–Hopf integral equation, which cannot be solved by inspection. Returning to (13), we can now write

$$\hat{\mathbf{x}}(t \mid t) = \int_a^t \overline{\mathbf{x}(t)\mathbf{v}'(s)}\mathbf{R}^{-1}(s)\mathbf{v}(s)ds. \tag{14}$$

This is the general formula for the *linear* least-squares estimate of $\mathbf{x}(t)$ from a white-noise process. (We may point out, in anticipation, that the *nonlinear* (nl) least-squares estimate is given by the remarkably similar formula

$$\hat{\mathbf{x}}_{\mathrm{nl}}(t \mid t) = \int_a^t \widehat{\mathbf{x}(t)\mathbf{v}'(s)}\mathbf{R}^{-1}(s)\mathbf{v}(s)ds \tag{15}$$

where

$$\widehat{\mathbf{x}(t)\mathbf{v}'(s)} = E[\mathbf{x}(t)\mathbf{v}'(s) \mid \mathbf{y}(\tau),\ a \le \tau < s]. \tag{16}$$

This result will be derived in Part III [27].

So far we have made no special assumptions on $\mathbf{x}(t)$. Kalman and Bucy [22] assumed that $\mathbf{x}(t)$ satisfies the differential equation

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{u}(t), \qquad t \ge a, \qquad \mathbf{x}(a) = \mathbf{x}_a \tag{17}$$

where $\mathbf{u}(\cdot)$ is white noise with intensity matrix $\mathbf{Q}(\cdot)$ and uncorrelated with the observation white noise $\mathbf{v}(\cdot)$, i.e.,

$$\overline{\mathbf{u}(t)\mathbf{u}'(s)} = \mathbf{Q}(t)\delta(t-s), \qquad \overline{\mathbf{u}(t)\mathbf{v}'(s)} \equiv 0 \tag{18}$$[4]

and the initial value $\mathbf{x}_a$ is a zero-mean random variable with variance $\mathbf{P}_a$ and uncorrelated with $\mathbf{u}(\cdot)$, i.e.,

$$\bar{\mathbf{x}}_a = 0, \quad \overline{\mathbf{x}_a\mathbf{x}_a'} = \mathbf{P}_a, \quad \overline{\mathbf{u}(s)\mathbf{x}_a'} \equiv 0, \quad a \le s < b. \tag{19}$$

To exploit this structure of $\mathbf{x}(t)$, we can differentiate the general estimate formula (14) to obtain

$$\begin{aligned}\dot{\hat{\mathbf{x}}}(t \mid t) &= \overline{\mathbf{x}(t)\mathbf{v}'(t)}\mathbf{R}^{-1}(t)\mathbf{v}(t) \\ &\quad + \left[\int_a^t \frac{d}{dt}\overline{\mathbf{x}(t)\mathbf{v}'(s)}\mathbf{R}^{-1}(s)\mathbf{v}(s)ds\right] \qquad (20)\\ &= \overline{\mathbf{x}(t)\mathbf{v}'(t)}\mathbf{R}^{-1}(t)\mathbf{v}(t) \\ &\quad + \left[\mathbf{F}(t)\int_a^t \overline{\mathbf{x}(t)\mathbf{v}'(s)}\mathbf{R}^{-1}(s)\mathbf{v}(s)ds\right. \\ &\quad \left. + \int_a^t \overline{\mathbf{u}(t)\mathbf{v}'(s)}\mathbf{R}^{-1}(s)\mathbf{v}(s)ds\right]. \qquad (21)\end{aligned}$$

Now the second term in (21) is equal to $\mathbf{F}(t)\hat{\mathbf{x}}(t \mid t)$ [cf. (14)], and thus but for the last term, (21) would be a differential equation for $\hat{\mathbf{x}}(t \mid t)$.

However, this last term will be zero if we assume that the white noise $\mathbf{u}(\cdot)$ that generates the signal process $\mathbf{x}(\cdot)$ is uncorrelated with the past observations $\mathbf{y}(\cdot)$ [and therefore with the equivalent observations $\mathbf{v}(\cdot)$]. That is, with the further assumption

$$\overline{\mathbf{u}(t)\mathbf{y}'(s)} \equiv 0, \qquad s < t \tag{22}$$

we shall have

$$\dot{\hat{\mathbf{x}}}(t \mid t) = \mathbf{F}(t)\hat{\mathbf{x}}(t \mid t) + \mathbf{K}(t)\mathbf{v}(t), \quad \mathbf{v}(t) = \mathbf{y}(t) - \hat{\mathbf{z}}(t \mid t) \tag{23}$$

where we have defined

$$\mathbf{K}(t) \triangleq \overline{\mathbf{x}(t)\mathbf{v}'(t)}\mathbf{R}^{-1}(t). \tag{24}$$

A block diagram for (23) is shown in Fig. 1(a) where the box yielding $\hat{\mathbf{z}}(t \mid t)$ will have a detailed structure similar to that for $\hat{\mathbf{x}}(t \mid t)$ if we assume that the $\mathbf{z}(t)$ obey a differential relation similar to (17) for $\mathbf{x}(t)$. We can be somewhat more explicit about $\hat{\mathbf{z}}(t \mid t)$ [and about the $\mathbf{K}(t)$ of (24)] if we assume some specific functional relationship between $\mathbf{z}(\cdot)$ and (past)[5] $\mathbf{x}(\cdot)$. The simplest is, of course, the linear relationship, used by Kalman and Bucy [22],

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) \tag{25}$$

which immediately yields (by linearity)

$$\hat{\mathbf{z}}(t \mid t) = \mathbf{H}(t)\hat{\mathbf{x}}(t \mid t). \tag{26}$$

This is very useful because now [the innovations $\mathbf{v}(t)$ can be obtained directly from $\hat{\mathbf{x}}(t \mid t)$ and $\hat{\mathbf{y}}(t)$] the estimate $\hat{\mathbf{x}}(t \mid t)$ can be realized by the *feedback* structure of Fig. 1(b).

The (*gain*) function $\mathbf{K}(t)$ can also be written in a simpler form under the assumption (25):

$$\begin{aligned}\mathbf{K}(t) &= \overline{\mathbf{x}(t)\mathbf{v}'(t)}\mathbf{R}^{-1}(t) \\ &= \overline{\mathbf{x}(t)[\tilde{\mathbf{x}}'(t \mid t)\mathbf{H}'(t) + \mathbf{v}'(t)]}\mathbf{R}^{-1}(t) \\ &= \overline{[\hat{\mathbf{x}}(t \mid t) + \tilde{\mathbf{x}}(t \mid t)]\tilde{\mathbf{x}}'(t \mid t)}\mathbf{H}'(t)\mathbf{R}^{-1}(t) + 0 \qquad (27)\\ &= 0 + \overline{\tilde{\mathbf{x}}(t \mid t)\tilde{\mathbf{x}}'(t \mid t)}\mathbf{H}'(t)\mathbf{R}^{-1}(t) \\ &= \mathbf{P}(t, t)\mathbf{H}'(t)\mathbf{R}^{-1}(t), \qquad \text{say} \qquad (28)[6]\end{aligned}$$

where

$\mathbf{P}(t, t)$ = the covariance function of the error in the estimate at time t.

It is easy to derive a differential equation for $\mathbf{P}(t, t)$ by first noting from (17) and (23) that $\tilde{\mathbf{x}}(t \mid t)$ obeys the differential equation

$$\begin{aligned}\dot{\tilde{\mathbf{x}}}(t \mid t) &= [\mathbf{F}(t) - \mathbf{K}(t)\mathbf{H}(t)]\tilde{\mathbf{x}}(t \mid t) \\ &\quad - \mathbf{K}(t)\mathbf{v}(t) + \mathbf{u}(t), \qquad \tilde{\mathbf{x}}(a \mid a) = \mathbf{x}_a. \qquad (29)\end{aligned}$$

Now applying a standard formula (Appendix I-B) we can show that $\mathbf{P}(t, t)$ satisfies the (nonlinear) matrix Riccati equation

$$\begin{aligned}\dot{\mathbf{P}}(t, t) &= \mathbf{F}(t)\mathbf{P}(t, t) + \mathbf{P}(t, t)\mathbf{F}'(t) - \mathbf{K}(t)\mathbf{R}(t)\mathbf{K}'(t) \\ &\quad + \mathbf{Q}(t), \qquad \mathbf{P}(a, a) = \mathbf{P}_a.\end{aligned} \tag{30}$$

[4] The assumption $\overline{\mathbf{u}(t)\mathbf{v}](s)} \equiv 0$ can be relaxed to $\overline{\mathbf{u}(t)\mathbf{v}'(s)} = \mathbf{C}(t)\delta(t-s)$; cf. (31) and (32).

[5] If $\mathbf{z}(\cdot)$ depended on future $\mathbf{x}(\cdot)$, we could not satisfy the condition (22).

[6] Note that (28) is true for general $\mathbf{x}(\cdot)$, not only those with the differential representation (17).

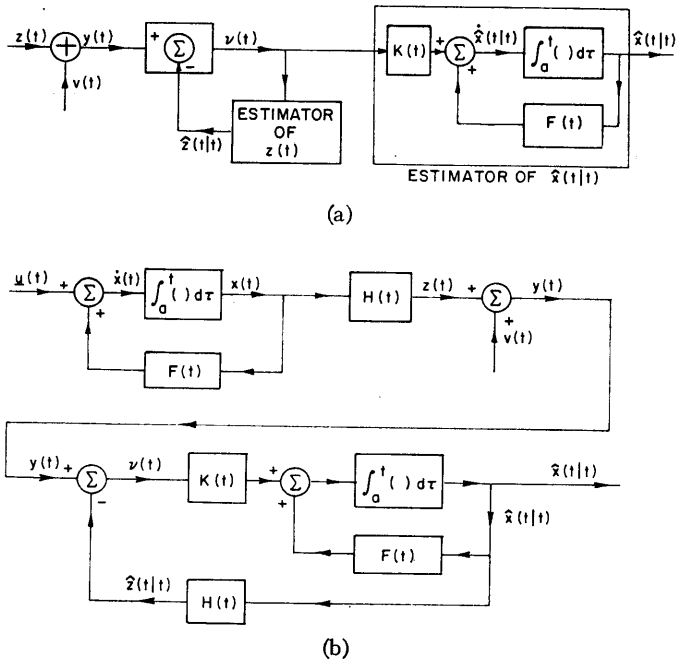


Fig. 1. (a) Filtered estimate of $x(t)$ from a related process $y(\tau)=z(\tau)+v(\tau)$, $a\leq\tau\leq t$. (b) The Kalman–Bucy filter; note that feedback of $\hat{x}(t|t)$ can be used to obtain $\nu(t)$ when $z(t)=H(t)x(t)$.

We now have obtained in formulas (23), (26), (28), and (30) the basic Kalman–Bucy formulas [for the problem defined by (1), (25), and (17)–(19)]. Our derivation is more direct than that of the original and reveals clearly the roles of the various assumptions in the Kalman-Bucy model. In Part III, we shall see the role that the corresponding assumptions play in the nonlinear problem. Our present proof also indicates some points at which the above arguments can be generalized. However, before doing this let us make a few supplementary remarks.

Correlated $u(\cdot)$ and $v(\cdot)$: We can, without violating the basic constraint (2) that $\overline{z(t)v'(s)}=0$, $s>t$, generalize the uncorrelatedness condition in (18) to

$$\overline{u(t)v'(s)} = C(t)\delta(t-s). \tag{31}$$

This will require minor changes in the above derivations,[7] which we shall leave for the reader's amusement. We shall only point out that finally the only change in the filter formulas [(23), (26), (28), (30)] will be that the gain function of (28) must be replaced by

$$\boldsymbol{K}(t) = [\boldsymbol{P}(t,t)'\boldsymbol{H}'(t) + \boldsymbol{C}(t)]\boldsymbol{R}^{-1}(t). \tag{32}$$

The Prediction Problem: Suppose we are to estimate $x(t+\Delta)$, $\Delta>0$, given observations $\nu(\cdot)$ up to t. Then, by the innovations technique, we readily find

$$\hat{x}(t+\Delta\,|\,t) = \int_a^t \overline{x(t+\Delta)\nu'(s)}\boldsymbol{R}^{-1}(s)\nu(s)ds. \tag{33}$$

[7] Notably that in (21) and (27) the terms that are zero will now be $(1/2)C(t)R^{-1}(t)$. The 1/2 arises from taking $\int_a^t\delta(t-s)\,ds=1/2$.

If $\boldsymbol{x}(\cdot)$ is a lumped process described by the model (17)–(19), then it is easy to see that

$$\hat{\boldsymbol{x}}(t+\Delta\,|\,t) = \boldsymbol{\Psi}(t+\Delta,t)\hat{\boldsymbol{x}}(t\,|\,t) \tag{34}$$

where $\boldsymbol{\Psi}(t,s)$ is the fundamental (or state-transition) matrix of the differential equation (17) of the process $\boldsymbol{x}(\cdot)$, i.e., $\boldsymbol{\Psi}(t,s)$ is the solution of

$$\frac{d}{dt}\boldsymbol{\Psi}(t,s) = \boldsymbol{F}(t)\boldsymbol{\Psi}(t,s), \qquad \boldsymbol{\Psi}(s,s)=\boldsymbol{I}. \tag{35}$$

The Steady-State Equation: In the preceding discussion, we have restricted the interval (a,t) to be finite. When the various matrices $\boldsymbol{F}(\cdot)$, $\boldsymbol{H}(\cdot)$, $\boldsymbol{Q}(\cdot)$, and $\boldsymbol{R}(\cdot)$, are time invariant, it is of interest to study the limiting behavior of the filter as the initial point a tends to $-\infty$. By a careful examination of the Riccati equation (30), Kalman and Bucy [22] have shown that when the model (17), (25) satisfies certain assumptions (stability, controllability, and observability, etc.), we can obtain a well-defined limiting solution by setting $\dot{\boldsymbol{P}}=0$ in (30) and using the non-negative[8] solution of the resulting algebraic equations in the filter formulas (23) and (27). When the process $\boldsymbol{x}(\cdot)$ has a rational spectral density, the above conditions are always met and the Kalman–Bucy solution reduces to the classical solution of Wiener. (The explicit equivalence has been shown by Leake [39].)

B. The Discrete-Time Problem

For discrete-time observations we will have similar results, with one rather trivial modification: the innovation process in the discrete-time case will have a different variance from that of the observation noise. Thus, let

$$\begin{aligned} &\boldsymbol{y}(k) = \boldsymbol{z}(k) + \boldsymbol{v}(k), && k=0,1,2,\cdots,\\ &\overline{\boldsymbol{v}(k)} = 0, && \overline{\boldsymbol{v}(k)\boldsymbol{v}'(l)} = \boldsymbol{R}(k)\delta_{kl} \end{aligned} \tag{36}$$

with $\{\boldsymbol{z}(k)\}$ a zero-mean finite-variance signal process. The innovation process will be defined by

$$\boldsymbol{\nu}(k) \triangleq \boldsymbol{y}(k) - \hat{\boldsymbol{z}}(k\,|\,k-1) \tag{37}$$

where

$$\begin{aligned} \hat{\boldsymbol{z}}(k\,|\,k-1) = {}& \text{the linear least-squares estimate of}\\ & \boldsymbol{z}(k) \text{ given } \{\boldsymbol{y}(l),\ 0\leq l\leq k-1)\}. \end{aligned} \tag{38}$$

Then it is easy to calculate (cf. Appendix II) that

$$\overline{\boldsymbol{\nu}(k)} = 0, \qquad \overline{\boldsymbol{\nu}(k)\boldsymbol{\nu}'(l)} = [\boldsymbol{P}_z(k)+\boldsymbol{R}(k)]\delta_{kl} \tag{39}$$

where

$$\begin{aligned} \boldsymbol{P}_z(k) = {}& \text{covariance matrix of the error in the}\\ & \text{estimate } \tilde{\boldsymbol{z}}(k\,|\,k-1)\\ = {}& \overline{[\boldsymbol{z}(k)-\hat{\boldsymbol{z}}(k\,|\,k-1)][\boldsymbol{z}(k)-\hat{\boldsymbol{z}}(k\,|\,k-1)]'}. \end{aligned} \tag{40}$$

Therefore, the innovation process is still (discrete-time)

[8] There are several solutions that are not non-negative definite.

white, but with a different variance. The estimation solution now proceeds essentially as in the continuous-time case; we shall rapidly outline the steps for a process $z(\cdot)$ of the form

$$\begin{aligned} z(k) &= H(k)x(k), \\ x(k+1) &= \Phi(k+1, k)x(k) + u(k), \\ \overline{u(k)u'(l)} &= Q(k)\delta_{kl}, \quad \overline{u(k)v'(l)} = C(k)\delta_{kl}. \end{aligned} \tag{41}$$

By the projection theorem, and assuming $[P_z(\cdot) + R(\cdot)]^{-1}$ exists,[9] we readily obtain the expression (42) for $\hat{x}(k+1 \mid k)$ in terms of the $v(l)$, $l \leq k$, which we can rearrange as

$$\hat{x}(k+1 \mid k) = \sum_0^k \overline{x(k+1)v'(l)}[P_z(l) + R(l)]^{-1}v(l) \tag{42}$$

$$= \sum_0^{k-1} \overline{x(k+1)v'(l)}[P_z(l) + R(l)]^{-1}v(l) + \overline{x(k+1)v'(k)}[P_z(k) + R(k)]^{-1}v(k) \tag{43}$$

$$= \Phi(k+1, k)\hat{x}(k \mid k-1) + K(k)v(k), \quad \text{say} \tag{44}$$

where we have defined

$$K(k) \triangleq \overline{x(k+1)v'(k)}[P_z(k) + R(k)]^{-1}. \tag{45}$$

Now we note that

$$\overline{x(k)v'(k)} = \overline{[\Phi(k+1, k)x(k) + u(k)][\tilde{x}'(k \mid k-1)H'(k) + v'(k)]} \tag{46}$$

$$= \Phi(k+1, k)\overline{x(k)\tilde{x}'(k \mid k-1)}H'(k) + C(k) = \Phi(k+1, k)P(k)H'(k) + C(k) \tag{47}$$

where

$$P(k) \triangleq \overline{\tilde{x}(k \mid k-1)\tilde{x}'(k \mid k-1)}. \tag{48}$$

Therefore, using

$$P_z(k) \triangleq \overline{\tilde{z}(k \mid k-1)\tilde{z}'(k \mid k-1)} = H(k)P(k)H'(k) \tag{49}$$

we can write $K(k)$ as

$$K(k) = \Phi(k+1, k)[P(k)H'(k) + C(k)] \cdot [H(k)P(k)H'(k) + R(k)]^{-1}. \tag{50}$$

Finally, with patience, we can derive a recursion relation for $P(k)$ which we quote without proof:

$$\begin{aligned} P(k+1) &= \Phi(k+1, k)A(k)\Phi'(k+1, k) + Q(k), \\ A(k) &= P(k) - K(k)[P_z(k) + R(k)]K'(k). \end{aligned} \tag{51}$$

Equations (44) and (51) define the discrete-time Kalman filter, first derived in a slightly less direct way (but still using the innovations) in Kalman [19]. Our method here is exactly parallel to the one we used in the continuous-time case.

[9] If not, we use the Moore-Penrose pseudo-inverse, but we shall not pursue this refinement here.

C. Some Generalizations

The crucial step in our derivation of the Kalman–Bucy formulas was the use of the assumptions

$$\dot{x}(t) = F(t)x(t) + u(t), \qquad \overline{u(t)v'(s)} \equiv 0$$

to write

$$\int_a^t \overline{\dot{x}(t)v'(s)}v(s)ds = F(t)\int_a^t \overline{x(t)v'(s)}v(s)ds = F(t)\hat{x}(t \mid t).$$

However, suppose we had

$$\dot{x}(t) = F(t)x(t-1) + u(t), \qquad \overline{u(t)y'(s)} \equiv 0. \tag{52}$$

Then we shall have

$$\int_a^t \overline{\dot{x}(t)v'(s)}v(s)ds = F(t)\int_a^t \overline{x(t-1)v'(s)}v(s)ds = F(t)\hat{x}(t-1 \mid t). \tag{53}$$

Kwakernaak [35] was apparently the first to point out this result. More generally, suppose

$$\begin{aligned} \dot{x}(t) &= \mathfrak{F} \circ x(\cdot) + u(t), \\ y(t) &= \mathfrak{H} \circ x(\cdot) + v(t), \qquad \overline{u(t)v'(t)} \equiv 0 \end{aligned} \tag{54}$$

where $\mathfrak{F} \circ x(\cdot)$ and $\mathfrak{H} \circ x(\cdot)$ denote some linear operation on the "past" values $\{x(s), a \leq s < t\}$ of the signal process. Then

$$\int_a^t \overline{\dot{x}(t)v'(s)}v(s)ds = \mathfrak{F} \circ \int_a^t \overline{x(\cdot)v'(s)}v(s)ds = \mathfrak{F} \circ \hat{x}(\cdot \mid t) \tag{55}$$

and the obvious analogs of the Kalman–Bucy formulas (23)–(30) are again easily obtained (of course, suitable attention has to be paid to the proper topologies, etc.). Some problems of this type have been noted by Balakrishnan and Lions [36] and Falb [35] who use essentially an operator–theoretic analog of the Kalman–Bucy derivation. They given some specific examples with $\mathfrak{F}$ being a partial differential operator. For a different illustration, we note that $\mathfrak{H}$ may be a random sampling operation, a case that was recently studied in a less direct manner by Chang [37]. General representations of the form (39) often arise in describing stochastic process by evolution equations in abstract spaces and, in fact, some nonlinear processes may be made linear by such representations. We shall not explore this point further in the present elementary paper.

However, it may be of some value to point out that if $x(\cdot)$ obeys a nonlinear equation

$$\dot{x}(t) = f(x(s), s \leq t, t) + u(t) \tag{56}$$

then the (linear) estimate $\hat{x}(t \mid t)$ obeys the equation

$$\dot{\hat{x}}(t \mid t) = \widehat{f(x(s), s \leq t, t)} + \overline{x(t)v'(t)}v(t) \tag{57}$$

where $\widehat{f(x(s), s \leq t, t)}$ is the best linear estimate of $f(x(\cdot), t)$ given $y(\tau)$, $a \leq \tau \leq t$. Such a problem was partially discussed by Chang [38].

Finally, we should make a brief comment about problems in which the additive observation noise is nonwhite. One solution is to apply a transformation that will whiten this noise and then use the Kalman–Bucy formulas. This method has been used by Bryson and Johansen [40]. However, a more powerful method is to whiten the whole observation process, the sum of the signal and the nonwhite noise; in other words, to obtain the innovations directly. This method is discussed in Geesey and Kailath [26]. It may be noted that the case of colored (finite-variance) noise plus white noise can be immediately treated by an obvious extension of Theorem 2—the observations can be whitened by subtracting out the estimates of the signal *and* the colored noise.

IV. Concluding Remarks

The main point of the innovations approach to statistical problems is that once we understand the basic probabilistic structure of the processes involved, many results can be obtained quite directly without resort to often more sophisticated (and analytical rather than probabilistic) tools like Wiener–Hopf techniques, Karhunen–Loève expansions, function space integrals, etc. In this paper we have illustrated this point for a class of nonstationary filtering problems.

In [25]–[31] applications are given for linear smoothing problems, nonlinear filtering and smoothing, covariance factorization, and detection problems.

Appendix I

A. The Projection Theorem

Formal proofs of the projection theorem are given in many textbooks. Here we shall make a few informal remarks that may aid in the understanding and application of the result. The projection theorem is probably quite familiar for linear (Hilbert) spaces of time function with inner product

$$\int_T u(t)v(t)dt \quad \text{or} \quad \int_T u(t)v(t)p(t)dt, \quad \text{where } p(t) \geq 0. \tag{58}$$

Thus, the linear least-squares approximation to an unknown function $u(\cdot)$ in terms of a given function $v(\cdot)$ is obtained by projecting $u(\cdot)$ on $v(\cdot)$ with the given inner product (58). For our applications, we need to work with Hilbert spaces of random variables, these being values of a stochastic process $z(t)$, for different time instants $t \in [a, b]$, or linear combinations of such random variables. Now random variables are also functions not of t, but of a probability sample-space variable, say $\omega \in \Omega$. The inner product is (very heuristically) $\int_\Omega u(\omega)v(\omega)p(\omega)d\omega$ where $p(\omega)\,d\omega$ is a probability, or, as it is usually written, $\overline{uv}$. As long as we remember that ω, the probability variable, should replace time, all our intuitive notions of Hilbert function spaces (which are essentially generalizations of n-dimensional Euclidean space) carry over to random variables. The orthogonality relations of the projection theorem have a geometric setting in this space of random variables. In this context, there is often some initial confusion because the variable t is also present in the discussion of stochastic processes. However, it is essential to remember that in the Hilbert space of random variables, the elements are not functions of time but functions of ω; the variable t serves only to index some of the elements of the Hilbert space.

B. Covariance Relations for Lumped Processes

Let a random process $\mathbf{x}(t)$ be obtained as the solution of the differential equation

$$\dot{\mathbf{x}}(t) = F(t)\mathbf{x}(t) + \mathbf{u}(t), \quad \mathbf{x}(a) = \mathbf{x}_a, \quad t > a \tag{59}$$

where (the zero means are assumed for notational convenience)

$$\overline{\mathbf{u}(t)} = 0, \qquad \overline{\mathbf{u}(t)\mathbf{u}'(s)} = \mathbf{Q}(t)\delta(t-s),$$
$$\bar{\mathbf{x}}_a, \overline{\mathbf{u}(t)\mathbf{x}_a'} = 0, \qquad t \geq a.$$

Then we can write

$$\mathbf{x}(t) = \mathbf{\Psi}(t, a)\mathbf{x}(a) + \int_a^t \mathbf{\Psi}(t, s)\mathbf{u}(s)ds$$

where $\mathbf{\Psi}(t, s)$ is the state-transition matrix defined as the (unique) solution of the equation

$$\frac{d\mathbf{\Psi}(t, s)}{dt} = F(t)\mathbf{\Psi}(t, s), \quad \mathbf{\Psi}(a, a) = I, \quad a \leq s \leq t. \tag{60}$$[10]

By direct computation, we obtain $\bar{\mathbf{x}}(t) \equiv 0$ and

$$\begin{aligned} \boldsymbol{R}_x(t, t) &\triangleq \overline{[\mathbf{x}(t) - \bar{\mathbf{x}}(t)][\mathbf{x}(t) - \bar{\mathbf{x}}(t)]'} \\ &= \mathbf{\Psi}(t, a)\boldsymbol{R}_a\mathbf{\Psi}'(t, a)' \\ &\quad + \int_a^t \mathbf{\Psi}(t, s)\mathbf{Q}(s)\mathbf{\Psi}'(t, s)ds. \end{aligned} \tag{61}$$

Differentiating both sides of (61) with respect to t and using (60), we obtain

$$\frac{d\boldsymbol{R}_x(t, t)}{dt} = F(t)\boldsymbol{R}_x(t, t) + \boldsymbol{R}_x(t, t)F'(t) + \mathbf{Q}(t), \qquad t \geq a \tag{62}$$

$$\boldsymbol{R}_x(a, a) = \boldsymbol{R}_a.$$

Furthermore, it follows by direct computation that

$$\boldsymbol{R}_x(t, s) \triangleq \overline{\mathbf{x}(t)\mathbf{x}'(s)}$$

$$= \overline{\left[\mathbf{\Psi}(t, s)\mathbf{x}(s) + \int_s^t \mathbf{\Psi}(t, \sigma)\mathbf{G}(\sigma)\mathbf{u}(\sigma)d\sigma\right]\mathbf{x}'(s)} \tag{63}$$

$$= \mathbf{\Psi}(t, s)\boldsymbol{R}_x(s, s) + 0 \qquad \text{for } t \geq s$$

$$= \boldsymbol{R}_x(t, t)\mathbf{\Psi}'(s, t) \qquad \text{for } s \geq t \tag{64}$$

where the last equation follows from the symmetry property $\boldsymbol{R}_x(t, s) = \boldsymbol{R}_x'(s, t)$. Equation (62), when applied to (29), yields the Riccati equation (30), as some

[10] When $F(\cdot)$ is time invariant, $\Psi(t,s) = e^{F(t-s)}$, $t \geq s$.

simple algebra will show. Equations (63) and (64) will be used for the smoothing problem in Part II. The above formulas are all well known.

Appendix II

The Innovation Process

If $y(\cdot)=z(\cdot)+v(\cdot)$, where $z(\cdot)$ is a second-order process and $v(\cdot)$ is white noise, we shall prove that the innovation process

$$\nu(t) = y(t) - \hat{z}(t \mid t), \qquad -\infty < a \leq t < b < \infty$$

is white with the same covariance as $v(\cdot)$, and that it is obtained from $y(\cdot)$ by a causal *invertible* linear operation. The first property follows easily by direct computation [and had been known for lumped process $z(\cdot)$]. The second property is more interesting and will be discussed first. (For simplicity, only the scalar case will be treated.)

A. The Relationship Between $y(\cdot)$ and $\nu(\cdot)$

Let $g_y(t, s)$ denote the optimum causal filter that operates on $\{y(s), s \leq t\}$ to give $\hat{z}(t \mid t)$, i.e.,

$$\hat{z}(t \mid t) = \int_a^t g_y(t, s)y(s)ds = \mathcal{G}_y y, \qquad \text{say} \qquad (65)$$

where $\mathcal{G}_y$ denotes the integral operator with kernel $g_y(t, s)$.[11]

To make (65) well defined we need to assume that (cf. Doob [9], sec. 9.2)

$$\int_a^t g_y^2(t, s)ds < \infty \qquad \text{for every } t \in (a, b). \qquad (66)$$

(If $g(t, \cdot)$ had delta functions in it, $\hat{z}(t \mid t)$ would have infinite variance.) From our assumption that $\int z^2(t)dt < \infty$, it can be shown that

$$\int_a^b \int_a^b g_y^2(t, s)dtds < \infty, \qquad (67)$$

a fact that will be useful presently. If we use I for the identity operator [the integral operator with kernel $\delta(t-s)$], then we can write, symbolically,

$$\nu = y - \hat{z} = y - \mathcal{G}_y y = (I - \mathcal{G}_y)y. \qquad (68)$$

The problem, then, is to show that $(I-\mathcal{G}_y)$ is a causally invertible operator. The causality of $\mathcal{G}_y$ does the trick here because $\mathcal{G}_y$ is then what is called a Volterra kernel and it can be proved (see, e.g., Smithies [41], p. 34) that when $\mathcal{G}_y$ has a square-integrable kernel, then $(1-\mathcal{G}_y)^{-1}$ exists and is given by the Neumann (geometric) series

$$(1 - \mathcal{G}_y)^{-1} = 1 + \mathcal{G}_y + \mathcal{G}_y^2 + \mathcal{G}_y^3 + \cdots \qquad (69)$$

where $\mathcal{G}_y^2 y = \mathcal{G}_y \mathcal{G}_y y$, and so on. The causality is obvious from (69).

In many applications, the signal process $z(\cdot)$ is continuous in the mean [which is equivalent to the continuity of the covariance function of $z(\cdot)$]. In this case, it can easily be shown that the kernel $g_y(t, s)$ is continuous in t and s (and, in this case, the arguments to establish (69) are even simpler (Riesz and Nagy [42], sec. 65).

B. The Process $\nu(\cdot)$ is White

We shall establish by direct calculation that

$$\overline{\nu(t)\nu(s)} = \overline{v(t)v(s)} \quad \text{where} \quad \nu(t) = y(t) - \hat{z}(t \mid t).$$

First consider $t>s$. Then

$$\begin{aligned} \overline{\nu(t)\nu(s)} &= \overline{[\tilde{z}(t \mid t) + v(t)][\tilde{z}(s \mid s) + v(s)]} \\ &= \overline{v(t)v(s)} + \overline{v(t)\tilde{z}(s \mid s)} \\ &\quad + \overline{\tilde{z}(t \mid t)\tilde{z}(s \mid s)} + \overline{\tilde{z}(t \mid t)v(s)}. \end{aligned} \qquad (70)$$

Now $\tilde{z}(s \mid s) = z(s) - \hat{z}(s \mid s)$ depends only on signal and noise up to time s. Since we have assumed that future noise is uncorrelated with past signal, the second term in (70) will be zero. Similarly, by the definition of $\tilde{z}(t \mid t)$, $\overline{\tilde{z}(t \mid t)\tilde{z}(s \mid s)} = \overline{\tilde{z}(t \mid t)z(s)} - 0$ for $t>s$. Therefore, we can write (70) as

$$\begin{aligned} \overline{\nu(t)\nu(s)} &= \overline{v(t)v(s)} + \overline{\tilde{z}(t \mid t)z(s)} + \overline{\tilde{z}(t \mid t)v(s)} \\ &= \overline{v(t)v(s)} + \overline{\tilde{z}(t \mid t)[z(s) + v(s)]} \\ &= \overline{v(t)v(s)} + \overline{\tilde{z}(t \mid t)y(s)} \\ &= \overline{v(t)v(s)} + 0, \qquad t > s. \end{aligned} \qquad (71)$$

A similar argument applies for $t<s$. Since $\overline{v(t)v(s)} = \delta(t-s) = 0$, $t \neq s$, we have $\overline{\nu(t)\nu(s)} = 0$, $t \neq s$. There remains only to examine the point $t=s$. Here we argue that $\overline{[\nu(t) - v(t)]^2} = \overline{\tilde{z}^2(t \mid t)} < \infty$, but $\overline{v^2(t)}$ is infinite (because $v(\cdot)$ is white), and therefore $\overline{\nu^2(t)}$ must be infinite (and we have just shown $\overline{\nu(t)\nu(s)} = 0$, $t \neq s$). This identifies $\nu(\cdot)$ as white noise. This argument seems shaky, but it is inevitable if we work with $v(\cdot)$ as an ordinary random process rather than as a generalized random process.

As long as we use the ordinary functional notation for $v(\cdot)$, all proofs, though they can be given in slightly different forms (especially with additional assumptions on $z(\cdot)$, e.g., that it is continuous in the mean or, more strongly, that it is a lumped process), must be essentially of the preceding form. A rigorous proof can be obtained by working with integrals of $v(\cdot)$ and $\nu(\cdot)$ (cf. [29]).

C. Discrete-Time Processes

Some insight is also shed on the preceding calculations by considering the discrete time

$$y(k) = z(k) + v(k), \qquad \overline{v(k)v(l)} = R(k)\delta_{kl},$$
$$\overline{v(k)z(l)} = 0, \qquad l \leq k.$$

In this case, the innovation process $\nu(\cdot)$ is defined as

$$\nu(k) = y(k) - \hat{z}(k \mid k-1) = \tilde{z}(k \mid k-1) + v(k) \qquad (72)$$

and, by arguments similar to those in (71), we obtain

[11] As an aside, we note that $\mathcal{G}_y$ can be regarded as an operator on L_2 (cf. Doob [9], sec. 9.2).

for $k>l$

$$\begin{aligned}\overline{\nu(k)\nu(l)} &= \overline{v(k)v(l)} + \overline{v(k)\hat{z}(l\,|\,l-1)} \\ &\quad + \overline{\hat{z}(k\,|\,k)-1)\hat{z}(l\,|\,l-1)} + \overline{\hat{z}(k\,|\,k-1)v(l)} \\ &= \overline{v(k)v(l)} + 0 + \overline{\hat{z}(k\,|\,k-1)[z(l)+v(l)]} \\ &= \overline{v(k)v(l)}.\end{aligned}$$

Similarly, we can prove the equality for $k<l$. For $k=l$, we have

$$\begin{aligned}\overline{\nu^2(k)} &= \overline{v^2(k)} + \overline{2v(k)\hat{z}(k\,|\,k-1)} + \overline{\hat{z}^2(k\,|\,k-1)} \\ &= \overline{v^2(k)} + \overline{\hat{z}^2(k\,|\,k-1)} = R(k) + P_z(k), \qquad \text{say.}\end{aligned}$$

Therefore

$$\overline{\nu(k)\nu(l)} = [R(k) + P_z(k)]\delta_{kl} \tag{73}$$

so that $\nu(\cdot)$, like $v(\cdot)$, is white but with a different variance. The continuous-time case can be approached by a limiting procedure in which $R(k)$ becomes indefinitely large while $P_z(k)$ remains finite, so that the variances of $\nu(\cdot)$ and $v(\cdot)$ are the same.

D. A Proof of the Equivalence of $\nu(\cdot)$ and $y(\cdot)$ Using the Kalman–Bucy Formulas

We noted in our discussion of Theorem 2 (cf. Remark 2) that the equivalence of $\nu(\cdot)$ and $y(\cdot)$ was obvious if the Kalman–Bucy result was assumed. The proof is trivial. Since $\nu(t)=y(t)-\hat{z}(t\,|\,t)$ and $\hat{z}(t\,|\,t)$ can be calculated from $y(s)$, $s\leq t$, $\nu(t)$ is completely determined by $y(s)$, $s\leq t$. Conversely, the Kalman–Bucy formula

$$\begin{aligned}\hat{x}(t\,|\,t) &= F(t)\hat{x}(t\,|\,t) + K(t)[y(t) - H(t)\hat{x}(t\,|\,t)], \\ \hat{x}(a\,|\,a) &= 0\end{aligned}$$

shows that $\hat{x}(t\,|\,t)$ is determined if $\{\nu(s),\ s\leq t\}$ is known, and then $y(t)$ can be obtained as $y(t)=H(t)\hat{x}(t\,|\,t)+\nu(t)$ since

$$\hat{z}(t\,|\,t)+\nu(t)=\hat{z}(t\,|\,t)+\tilde{z}(t\,|\,t)+v(t)=z(t)+v(t).$$

Therefore, $\nu(\cdot)$ and $y(\cdot)$ can each be obtained from the other by causal operations. This argument is due to R. Geesey.

Of course, the deeper result is that this fact is true without restriction to lumped processes and can indeed be used, as we have shown, to give a simple proof of the special formulas for lumped processes.

Acknowledgment

The author thanks R. Geesey, B. Gopinath, and P. Frost, former students at Stanford University, for many stimulating, enjoyable, and instructive conversations on various aspects of the innovations concept and, in particular, for teaching him various aspects of modern control theory.

References

[1] A. N. Kolmogorov, "Interpolation and extrapolation of stationary random sequences," *Bull. Acad. Sci. USSR, Math. Ser.* vol. 5, 1941. A translation has been published by the RAND Corp., Santa Monica, Calif., as Memo. RM-3090-PR.

[2] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications.* New York: Wiley, 1949. Originally issued as a classified report by M.I.T. Radiation Lab., Cambridge, Mass., February 1942.

[3] H. Wold, *A Study in the Analysis of Stationary Time Series.* Uppsala, Sweden: Almqvist & Wiksell, 1938.

[4] P. Whittle, *Prediction and Regulation.* Princeton, N. J.: Van Nostrand, 1963.

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

[6] 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, April 1950.

[7] R. B. Blackman, H. W. Bode, and C. E. Shannon, "Data smoothing and prediction in fire-control systems," Research and Development Board, Washington, D.C., August 1948.

[8] 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.

[9] J. L. Doob, *Stochastic Processes.* New York: Wiley, 1953.

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

[11] V. S. Pugachev, *Theory of Random Functions and Its Applications in Automatic Control.* Moscow: Goztekhizdat, 1960.

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

[13] A. M. Yaglom, *Theory of Stationary Random Functions*, R. A. Silverman, transl. Englewood Cliffs, N. J.: Prentice-Hall, 1966.

[14] R. Deutsch, *Estimation Theory.* Englewood Cliffs, N. J.: Prentice-Hall, 1966.

[15] P. B. Liebelt, *An Introduction to Optimal Estimation Theory.* Reading, Mass.: Addison-Wesley, 1967.

[16] A. V. Balakrishnan, "Filtering and prediction theory," in *Lectures on Communication Theory*, A. V. Balakrishnan, Ed. New York: McGraw-Hill, 1968.

[17] A. E. Bryson and Y. C. Ho, *Optimal Programming, Estimation and Control.* New York: Blaisdell, 1968.

[18] P. Swerling, "First-order error propagation in a stagewise smoothing procedure for satellite observations," *J. Astronautical Sci.*, vol. 6, no. 3, pp. 46–52, Autumn 1959. See also "A proposed stagewise differential correction procedure for satellite tracking and prediction," RAND Corp., Santa Monica, Calif., Rept. P-1292, January 1958.

[19] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Engrg.*, vol. 82, pp. 34–45, March 1960.

[20] ——, "New methods in Wiener filtering theory," *Proc. 1st Symp. on Engrg. Applications of Random Function Theory and Probability*, J. L. Bogdanoff and F. Kozin, Eds. New York: Wiley, 1963.

[21] R. S. Bucy, "Optimum finite-time filters for a special nonstationary class of inputs," Johns Hopkins University, Appl. Phys. Lab., Baltimore, Md., Internal Memo. BBD-600, 1959.

[22] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Trans. ASME, J. Basic Engrg.*, ser. D vol. 83, pp. 95–107, December 1961.

[23] W. M. Wonham, "Lecture notes on stochastic optimal control," Div. of Appl. Math., Brown University, Providence, R.I., Rept. 67-1.

[24] A. J. F. Siegert, "A systematic approach to a class of problems in the theory of noise and other random phenomena," Pt. 2 and 3, *IRE Trans. Information Theory*, vol. IT-3, pp. 38–43, March 1957; vol. IT-4, pp. 4–14, March 1958.

[25] T. Kailath and P. Frost, "An innovations approach to least-squares estimation—Part II: Linear smoothing in additive white noise," this issue, page 655.

[26] R. Geesey and T. Kailath, "An innovations approach to least-squares estimation—Part III: Estimation in colored noise" (to be published).

[27] ——, "An innovations approach to least-squares estimation—part IV: Nonlinear filtering and smoothing in white Gaussian noise" (to be published).

[28] P. A. Frost, "Estimation in continuous-time nonlinear systems," Ph.D. dissertation, Dept. of Elec. Engrg., Stanford University, Stanford, Calif., June 1968.

[29] ——, "A general likelihood ratio formula for random signals in Gaussian noise," *IEEE Trans. Information Theory*, to appear, 1969.

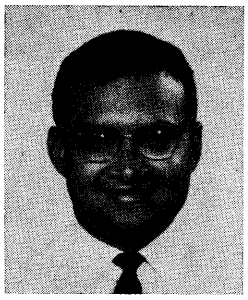
[30] T. Kailath, "An RKHS approach to detection and estimation—Part III: More on gaussian detection," to be submitted to *IEEE Trans. Information Theory.*

[31] T. Kailath and R. Geesey, "Covariance factoriztion—An explication via examples," *Proc. 2nd Asilomar Conference on Circuits and Systems*, Monterey, Calif., November 1968.

[32] H. J. Kushner, *Stochastic Stability and Control.* New York: Academic Press, 1967.

[33] L. D. Collins, "Realizable whitening filters and state-variable realizations," *Proc. IEEE (Letters)*, vol. 56, pp. 100–101, January 1968.

[34] B. D. O. Anderson and J. B. Moore, "Whitening filters: A state-space viewpoint," Dept. of Elec. Engrg., University of Newcastle, Australia, Tech. Rept. EE 6707, August 1967. Also see *Proc. JACC*, (Michigan). 1968.
[35] H. Kwakernaak, "Optimal filtering in linear systems with time delays," *IEEE Trans. Automatic Control*, vol. AC-12, pp. 169–173, April 1967.
[36] P. Falb, "Kalman-Bucy filtering in Hilbert space," *Information and Control*, vol. 11, no. 1, pp. 102–137, August–September 1967.
[37] A. V. Balakrishnan and J. L. Lions, "State estimation for infinite-dimensional systems," *J. Computer and System Sciences*, vol. 1, no. 4, pp. 391–403, December 1967.
[38] S. S. L. Chang, "Optimum filtering and control of randomly sampled systems," *IEEE Trans. Automatic Control*, vol. AC-12, pp. 537–546, October 1967.
[39] R. J. Leake, "Duality condition established in the frequency domain," *IEEE Trans. Information Theory (Correspondence)*, vol. IT-11, p. 461, July 1965.
[40] A. E. Bryson, Jr., and D. E. Johansen, "Linear filtering for time-varying systems using measurements containing colored noise," *IEEE Trans. Automatic Control*, vol. AC-10, pp. 4–10, January 1965.
[41] F. Smithies, *Integral Equations*. London: Cambridge University Press, 1958.
[42] F. Riesz and B. S. Nagy, *Functional Analysis*. New York: Ungar, 1955.



Thomas Kailath (S'57–M'62) was born in Poona, India, on June 7, 1935. He obtained the Bachelor's degree in telecommunications engineering at the University of Poona, Poona, India, in 1956, and the S.M. and Sc.D. degrees at the Massachusetts Institute of Technology, Cambridge, in 1959 and 1961, respectively.

He worked at the Jet Propulsion Laboratories, Pasadena, Calif., until 1963, and since then has been at Stanford University, Stanford, Calif., where he is now Professor of Electrical Engineering. He was a visiting scholar in the Department of Electrical Engineering of the University of California, Berkeley, from January to June, 1963. His research interests are in communication through time-variant channels, feedback communication systems, continuous-time detection and estimation problems, and the analysis and structure of stochastic systems. He was coauthor of a paper on feedback systems that received the 1967 Information Theory Group Award. He is Consulting Editor for a Prentice-Hall series on information theory.

Dr. Kailath is a member of SIAM, the Institute of Mathematical Statistics, URSI, and Sigma Xi.

An Innovations Approach to Least-Squares Estimation Part II: Linear Smoothing in Additive White Noise

THOMAS KAILATH, MEMBER, IEEE, AND PAUL FROST, MEMBER, IEEE

***Abstract*—The innovations method of Part I is used to obtain, in a simple way, a general formula for the smoothed (or noncausal) estimation of a second-order process in white noise. The smoothing solution is shown to be completely determined by the results for the (causal) filtering problem. When the signal is a lumped process, differential equations for the smoothed estimate can easily be derived from the general formula. In several cases, both the derivations and the forms of the solution are significantly simpler than those given in the literature.**

I. Introduction

IN THIS PAPER, we apply the innovations technique of Part I[1] to solve the smoothing problem. We shall show that the smoothing solution is completely determined in a simple way by the optimum causal filter and its adjoint. This result is valid for a general second-order (finite-variance) signal process in white noise, without restriction to lumped signal processes. For lumped processes, recursive solutions are available for the filtered estimate (Part I), and from these similar solutions can easily be found for the smoothed estimate. In the literature, recursive solutions to the smoothing problem have generally been obtained rather laboriously and often in less convenient form than ours (cf. Section III).

The problem we shall begin with is the following. We are given observations

$$y(t) = H(t)x(t) + v(t), \qquad a \leq t < b \tag{1}$$

where

$$\overline{v(t)} = 0, \quad \overline{v(t)v'(s)} = R(t)\delta(t-s), \quad R(t) > 0 \tag{2}$$

$$\overline{x(t)} = 0, \quad \overline{x(t)x'(t)} < \infty, \quad \overline{x(t)v'(s)} \equiv 0, \quad s > t. \tag{3}$$

It is required to find the linear least-squares smoothed estimate

$$\begin{aligned} x(t \mid b) = {} & \text{the linear function of the data} \\ & \{y(s), a \leq s < b\} \text{ that minimizes} \\ & \overline{[x(t) - \hat{x}(t \mid b)]'\,[x(t) - \hat{x}(t \mid b)]}. \end{aligned} \tag{4}$$

Manuscript received January 31, 1968. This work was supported by the Applied Mathematics Division of the Air Force Office of Scientific Research under Contract AF 49(638)1517, by the Air Force Avionics Laboratory under Contract F 33615-67-C-1245, and by the Joint Services Electronics Program at Stanford University, Stanford, Calif., under Contract Nonr 225(83).

The authors are with Stanford University, Stanford, Calif.

[1] This issue, page 646.