A Tutorial Introduction to Estimation and Filtering

IAN B. RHODES, MEMBER, IEEE

Abstract—In this tutorial paper the basic principles of least squares estimation are introduced and applied to the solution of some filtering, prediction, and smoothing problems involving stochastic linear dynamic systems. In particular, the paper includes derivations of the discrete-time and continuous-time Kalman filters and their prediction and smoothing counterparts, with remarks on the modifications that are necessary if the noise processes are colored and correlated. The examination of these state estimation problems is preceded by a derivation of both the unconstrained and the linear least squares estimator of one random vector in terms of another, and an examination of the properties of each, with particular attention to the case of jointly Gaussian vectors. The paper concludes with a discussion of the duality between least squares estimation problems and least squares optimal control problems.

I. Introduction

THIS paper contains a tutorial introduction to the basic principles of least squares estimation and their application to the solution of some state-estimation problems associated with finite-dimensional linear dynamic systems operating in a stochastic environment. The exposition begins with the problem of estimating one random variable or vector in terms of another, and proceeds through a derivation of the discrete-time Kalman filter and predictor to a derivation of the continuous-time Kalman–Bucy filter and predictor, and an examination of the continuous-time smoothing problem. The paper concludes with a discussion of the duality between least squares estimation and least squares optimal control problems.

The development in each section draws nontrivially on that in preceding sections and on the following assumed prerequisites.

1) Familiarity with the elements of probability theory through the concept of jointly distributed random variables and random vectors described by their joint probability density function, and the associated means, covariances, and conditional expectations. For a discussion of these topics, see, for example, [1].

2) Beginning with Section VII, an exposure to the state-space description of linear dynamic systems, the dual concepts of controllability and observability, and the least squares ("linear-quadratic") regulator problem in both finite and infinite time. For an exposition of these topics the reader is referred to the appropriate papers in this issue or, for example, to [2].

In Section II, we examine the least squares estimation of one random vector in terms of another. Some important properties of this estimator are derived in Section III, while in Section IV we summarize the properties of Gaussian random vectors and discuss the least squares estimation of one jointly Gaussian random vector in terms of another. Section V is devoted to the *linear* least squares estimation of one random vector in terms of another, and Section VI contains the derivations of a number of properties of this linear least squares estimator that are important both in their own right and for the straightforward inductive derivation, in Section VII, of the Kalman filter and predictor for estimating the state of a noise-corrupted discrete-time linear dynamic system in terms of its noisy output measurements. These same properties provide the background for the introduction of the so-called innovations sequence (or process) and the inclusion of an alternative derivation of the discrete-time Kalman filter and predictor. The continuous-time filtering, prediction, and smoothing problems are then examined from an innovations viewpoint in Section VIII. Finally, Section IX contains a discussion of the duality between least squares estimation problems and least squares control problems, and an examination of the steady-state Kalman–Bucy filter.

In the first six sections we distinguish between random vectors and their sample values by denoting the former with capital letters X, Y, and Z and the latter by the corresponding lower case letter x, y, or z. In order to conform with standard engineering usage, we discontinue making this distinction in Sections VII through IX where the lower case letters x, y, z, etc., are used to denote both random vectors and their sample values. Unless specifically indicated to the contrary, the upper case letters A, B, etc. denote matrices, while the lower case letters a, b, etc. denote (nonrandom) vectors.

II. Least Squares Estimation of One Random Vector in Terms of Another

Consider two jointly distributed random vectors X and Y, and suppose that in a particular sample observation we measure the value of the random vector Y to be (the m-vector) y. It is not unreasonable to expect that the knowledge that Y has value y will convey in general some information about the corresponding (but unmeasured) sample value x of the random vector X. In particular, we expect that in general the information that Y has value y will change any *a priori* guess or estimate we might have made about the value of X and that our degree of uncertainty about X will have been decreased. To be more specific, it is natural to ask the following question. Given the information that Y has the value y, what is the best

Manuscript received July 19, 1971. Paper recommended by D. G. Luenberger, Associate Guest Editor. The preparation of this paper was supported by the Air Force Office of Scientific Research through Themis Grant F44620-69-C-0116.

The author is with the Control Systems Science and Engineering Laboratory, Washington University, St. Louis, Mo.

estimate $\hat{x}$ of the corresponding value of the random vector X? The sense in which the estimate $\hat{x}$ is to be "best" must, of course, be defined. While several natural criteria present themselves, we restrict attention here to the case where we define the best estimate to be the n-vector $\hat{x}$ that minimizes over all n-vectors z the conditional expectation

$$E\{\|X - z\|^2|Y = y\} = E\{[X - z]'[X - z]|Y = y\}$$

of the norm-squared estimation error given that Y has value y. In other words, we adopt as our measure of uncertainty about X the mean-squared estimation error given that Y has value y, and we choose as the best estimate the n-vector $\hat{x}$ that minimizes this measure of uncertainty. For obvious reasons, this estimate is known variously as the least squares estimate, the least mean squares estimate, the minimum mean-square-error estimate, or the minimum variance estimate.

We now summarize and formalize these ideas in the following problem statement.

Problem 1 (Least Squares Estimate): Consider two jointly distributed random vectors X and Y with respective dimensions n and m and with joint probability density $f_{X,Y}(\cdot,\cdot)$, and suppose that in a particular sample observation the value of Y is measured to be y. Find the estimate $\hat{x}$ of the corresponding sample value x of the random vector X that is best in the sense that $\hat{x}$ minimizes over all n-vectors z the conditional expectation $E\{\|X - z\|^2|Y = y\}$ of the norm-squared estimation error given that Y has value y.

Proposition 1: The solution to Problem 1 is given for all y with $f_Y(y) > 0$ by the conditional expectation

$$\hat{x} = E\{X|Y = y\} = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx$$
$$= \int_{-\infty}^{\infty} \frac{x f_{X,Y}(x,y)}{f_Y(y)}\, dx \quad (1)$$

of X given that Y has value y. The corresponding minimum mean-square error is the conditional covariance of X given that Y has value y.

Proof: Expanding the optimality criterion, using the linearity of the expectation, and then "completing the square" we obtain

$$\begin{aligned} E\{\|X - z\|^2|Y = y\} &= E\{X'X - 2z'X + z'z|Y = y\} \\ &= E\{X'X|Y = y\} \\ &\quad - 2z'E\{X|Y = y\} + z'z \\ &= E\{\|z - E\{X|Y = y\}\|^2\} \\ &\quad + E\{\|X\|^2|Y = y\} \\ &\quad - \|E\{X|Y = y\}\|^2. \end{aligned} \quad (2)$$

The only term on the right side of (2) involving z is the first, and this is uniquely minimized by setting

$$z = \hat{x} = E\{X|Y = y\}. \quad (3)$$

The corresponding minimum value of (2) is then

$$\begin{aligned} E\{\|X - \hat{x}\|^2|Y = y\} &= E\{\|X\|^2|Y = y\} - \|E\{X|Y = y\}\|^2 \\ &= E\{\|X\|^2|Y = y\} - \|\hat{x}\|^2 \end{aligned} \quad (4)$$

which is the conditional variance of X given that $Y = y$. Alternatively, the same conclusion can be obtained by first differentiating the second line of (2) with respect to z and setting the result equal to the zero vector, and then noting that the second derivative of (2) with respect to z is twice the $n \times n$ unit matrix, so that the local extremum so obtained is indeed a (global) minimum. Q.E.D.

It is important to observe that the essence of Problem 1 is that the value y of the random vector Y is given and we seek the n-vector $\hat{x}$ that is the best estimate of the value of the random vector X. It is clear that $\hat{x}$ will depend in general on the given m-vector y. Conceptually, this procedure could be repeated for every value y of Y for which $f_Y(y) > 0$ to yield, in principle, a graph of the corresponding best estimate $\hat{x}$ in terms of the value y of Y. This graph may, of course, be interpreted as defining a function, which we denote $\hat{X}$, of the random vector Y. Since a function of a random vector is itself a random vector, it follows that $\hat{X}$ is a random vector—it is the random vector defined for all y with $f_Y(y) > 0$ by

$$\hat{X}(y) \triangleq \hat{x} = E\{X|Y = y\}. \quad (5a)$$

In other words, $\hat{X}$ is the random vector

$$\hat{X} = E\{X|Y\}. \quad (5b)$$

We can view $\hat{X}$ as an operator (or "black box") that accepts sample observations y on the random vector Y and produces the corresponding least squares estimate

$$\hat{x} = \hat{X}(y) = E\{X|Y = y\}$$

of X given that Y has value y. We call $\hat{X}$ the *least squares estimator* of X in terms of Y. If g is any function mapping R^m into R^n, then it is clear from the way we constructed $\hat{X}$ that for all y with $f_Y(y) > 0$ we have

$$E\{\|X - \hat{X}(Y)\|^2|Y = y\} \leqq E\{\|X - g(Y)\|^2|Y = y\}.$$

Taking expectations of both sides and invoking the identity[1] [1]

$$E_Y[E\{\|X - g(Y)\|^2|Y\}] = E\{\|X - g(Y)\|^2\} \quad (6)$$

we then have

$$E\{\|X - \hat{X}(Y)\|^2\} \leqq E\{\|X - g(Y)\|^2\}. \quad (7)$$

Thus the estimator $\hat{X}$ constructed by solving Problem 1 for

[1] This identity is easily proven by noting that the left side is just

$$\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \|x - g(y)\|^2 f_{X|Y}(x|y)\, dx\right] f_Y(y)\, dy$$

which, on writing $f_{X|Y}(x|y) f_Y(y) = f_{X,Y}(x, y)$ (Bayes' rule), reduces immediately to the right side.

each y with $f_Y(y) > 0$ is also the solution to the problem where we seek at the outset the estimator $\hat{X}$ that minimizes the unconditioned expectation $E\{\|X - g(Y)\|^2\}$ over all functions $g: R^m \to R^n$. In other words, we have the following problem and its solution.

Problem 1' (Least Squares Estimator): Consider two jointly distributed random vectors X and Y with joint probability density function $f_{X,Y}(\cdot,\cdot)$. Find the estimator $\hat{X}$ of X in terms of Y that is best in the sense that $\hat{X}$ minimizes $E\{\|X - g(Y)\|^2\}$ over all functions g mapping R^m into R^n.

Proposition 1': The least squares estimator $\hat{X}$ of X in terms of Y in the sense of Problem 1' is the conditional expectation

$$\hat{X} = E\{X|Y\}$$

of X given Y, and the corresponding minimum mean-square error is the conditional variance $E\{\|X - E\{X|Y\}\|^2\}$.

It is important to note that in Problem 1 we seek an *estimate*, viz., the vector $\hat{x}$ that minimizes over all n-vectors z the conditional expectation $E\{\|X - z\|^2|Y = y\}$ given that Y has value y, whereas in Problem 1' we seek an *estimator*, viz., the *function* $\hat{X}$ that minimizes the (unconditioned) expectation $E\{\|X - g(Y)\|^2\}$ over all functions g mapping R^m into R^n. As long as the functions g are unconstrained in any way (such as being continuous or linear) we can construct the solution to Problem 1' by solving Problem 1 for each y with $f_Y(y) > 0$, as we have done above. If, on the other hand, the functions g over which we seek a solution to Problem 1' are constrained to be, for example, linear or continuous, then this approach is no longer valid (unless, of course, the unconstrained least squares estimator turns out to have the desired property of linearity or continuity), and an alternative approach must be adopted. Depending on the joint probability density function involved, it may well turn out that in the constrained case there may be some functions g and values y for which

$$E\{\|X - \hat{X}(y)\|^2|Y = y\} > E\{\|X - g(y)\|^2|Y = y\}$$

but $\hat{X}$ is still the best "on the average" in the sense that (7) holds for all functions g satisfying the required constraints.

III. Properties of the Least Squares Estimator

The estimator of X in terms of Y that is best in the least squares sense of Problem 1' has a number of properties that are important both in their own right and in theoretical developments involving the least squares estimator such as the separation theorem [3]. Most of these properties are direct consequences of the fact that the least squares estimator of X in terms of Y is the conditional expectation $E\{X|Y\}$. We begin with three trivial properties.

Proposition 2a (Properties of the Least Squares Estimator): The least squares estimator $\hat{X} = E\{X|Y\}$ has the following properties.

1) It is linear, i.e., for any deterministic matrix A and deterministic vector b with the appropriate dimensions

$$\widehat{(AX + b)} = E\{AX + b|Y\} = AE\{X|Y\} + b = A\hat{X} + b$$

and if X and Z are random vectors with the same dimension

$$\widehat{(X + Z)} = E\{X + Z|Y\} = E\{X|Y\} + E\{Z|Y\} = \hat{X} + \hat{Z}.$$

2) It is unbiased in the sense that

$$E\{X - \hat{X}\} = E\{X\} - \underset{Y}{E}\{E[X|Y]\} = E\{X\} - E\{X\} = 0.$$

In fact, we have the stronger statement

$$E\{X - \hat{X}|Y\} = E\{X|Y\} - E[E\{X|Y\}|Y] = E\{X|Y\} - E\{X|Y\} = 0.$$

3) For any nonnegative matrix F, $\hat{X} = E\{X|Y\}$ minimizes $E\{[X - g(Y)]'F[X - g(Y)]\}$ over all functions $g: R^m \to R^n$, and $\hat{x} = E\{X|Y = y\}$ minimizes $E\{[X - z]'F[X - z]|Y = y\}$ over all n-vectors z.

Proof: The proofs of 1 and 2 are included in the statements of these properties. Property 3 follows by a direct modification of the proof of Proposition 1. If F is positive definite the proof is unchanged if $\|q\|^2$ is interpreted to mean $q'Fq$ and $z'q$ is replaced by $z'Fq$; if F is nonnegative definite the same identifications may be made but $\|q\| = [q'Fq]^{1/2}$ is in this case only a seminorm, and while $\hat{x} = E\{X|Y = y\}$ minimizes the first term on the right side of (2), it does not do so uniquely.

The following property of the least squares estimator is somewhat less trivial and, like the above three properties, has far reaching implications in estimation theory.

Proposition 2b: The estimation error $\tilde{X} \triangleq X - \hat{X}$ in the least squares estimator $\hat{X} = E\{X|Y\}$ is uncorrelated with any function g of the random vector Y, i.e.,

$$E\{g(Y)\tilde{X}'\} = 0 \tag{8}$$

and, in fact,

$$E\{\overset{n\times n}{g(Y)\tilde{X}'}|Y\} = 0. \tag{9}$$

Proof: For every value y of Y we have

$$\begin{aligned} E\{g(Y)\tilde{X}'|Y = y\} &= E\{g(y)[X - \hat{X}(y)]'|Y = y\} \\ &= g(y)E\{X' - \hat{X}'(y)|Y = y\} \\ &= g(y)[\hat{X}'(y) - \hat{X}'(y)] = 0 \end{aligned} \tag{10}$$

which is (9). Equation (8) then follows by taking expectations of both sides of (10) over Y and using the identity $\underset{Y}{E}[E\{g(Y)\tilde{X}'|Y\}] = E\{g(Y)\tilde{X}'\}$ [cf., (6)]. Q.E.D.

IV. Least Squares Estimation of Gaussian Random Vectors

Gaussian random vectors play a major role in probability theory and system theory. Their importance stems largely from two facts: first, they possess many distinctive mathematical properties and, second, the Gaussian distribution bears close resemblance to the probability laws

$$\underset{r\times 1}{Z} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

of many physical random phenomena. We summarize here the properties of Gaussian random vectors that are of greatest importance in estimation theory, leaving it to the reader to consult, e.g., [1], [4] for a detailed discussion of the Gaussian distribution.

Definition 1: An r-dimensional random vector $\boldsymbol{Z}$ is said to be *Gaussian* (or normal) with parameters $\boldsymbol{m}$ (an r-vector) and $\boldsymbol{\Sigma}$ (an $r \times r$ positive definite matrix) if its probability density function $f_Z(\cdot)$ is given for all $z \in R^r$ by

$$f_Z(z) = (2\pi)^{-r/2}|\boldsymbol{\Sigma}|^{-1/2} \exp[-\tfrac{1}{2}(z - \boldsymbol{m})'\boldsymbol{\Sigma}^{-1}(z - \boldsymbol{m})] \tag{11}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. The corresponding characteristic function of $\boldsymbol{Z}$ is

$$\varphi_Z(\boldsymbol{v}) = E\{\exp[j\boldsymbol{v}'\boldsymbol{Z}]\} = \exp[j\boldsymbol{v}'\boldsymbol{m} - \tfrac{1}{2}\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v}] \tag{12}$$

and this constitutes an alternative definition of a Gaussian random vector with parameters $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$. In fact, the definition of a Gaussian random vector as one whose characteristic function is given by (12) is more general because it includes the possibility that $\boldsymbol{Z}$ may be degenerate and have its entire density concentrated on a proper subspace of R^r, in which case $|\boldsymbol{\Sigma}| = 0$ and $\boldsymbol{\Sigma}$ is nonnegative definite but not positive definite and not invertible. In any case, we henceforth adopt the shorthand notation $N(\boldsymbol{m}, \boldsymbol{\Sigma})$ for the Gaussian (or normal) distribution with parameters $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$.

Using the well-known [1], [4] properties of the characteristic function to compute the moments of $\boldsymbol{Z}$ we have, in particular,

$$E\{\boldsymbol{Z}\} = \frac{1}{j} \cdot \frac{d\varphi}{d\boldsymbol{v}}(0) = \underset{}{\overset{r\times 1}{\boldsymbol{m}}}$$

$$\text{cov}[\boldsymbol{Z}, \boldsymbol{Z}] = \frac{1}{j^2} \cdot \frac{d^2\varphi}{d\boldsymbol{v}^2}(0) - \boldsymbol{m}\boldsymbol{m}' = \overset{r\times r}{\boldsymbol{\Sigma}}.$$

Thus the parameters $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ of the Gaussian probability distribution (11) or (12) are, respectively, the mean and the covariance of $\boldsymbol{Z}$. It is important to note that the probability density function of a Gaussian random vector is therefore completely specified by a knowledge of its mean and covariance. The importance of Gaussian random vectors in estimation and control theory is due largely to this fact and to the following facts.

1) Uncorrelated jointly Gaussian random vectors are independent.

2) Linear functions of Gaussian random vectors are themselves Gaussian random vectors.

3) In particular, sums of jointly Gaussian random vectors are Gaussian random vectors.

4) The conditional expectation of one jointly Gaussian random vector given another is a Gaussian random vector that is a linear function of the conditioning vector.

In particular, let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be jointly distributed random vectors with respective dimensions n and m whose composite vector $\boldsymbol{Z} = [\boldsymbol{X}', \boldsymbol{Y}']'$ is $N(\boldsymbol{m}, \boldsymbol{\Sigma})$ with mean

$r = n + m$

$$E[\boldsymbol{Z}] = \underset{r\times 1}{\boldsymbol{m}} = \overset{m\times 1}{\begin{bmatrix} \boldsymbol{m}_X \\ \cdots \\ \boldsymbol{m}_Y \end{bmatrix}} = \underset{m\times 1}{\begin{bmatrix} E[\boldsymbol{X}] \\ \cdots \\ E[\boldsymbol{Y}] \end{bmatrix}} \tag{13a}$$

and covariance

$$\text{cov}[\boldsymbol{Z}, \boldsymbol{Z}] = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \vdots & \boldsymbol{\Sigma}_{XY} \\ \cdots & & \cdots \\ \underset{n\times n}{\boldsymbol{\Sigma}_{YX}} & \vdots & \underset{n\times m}{\boldsymbol{\Sigma}_{YY}} \end{bmatrix} \qquad \Sigma_{xy} = (\Sigma_{yx})'$$

$$= \begin{bmatrix} \text{cov}[\boldsymbol{X}, \boldsymbol{X}] & \vdots & \text{cov}[\boldsymbol{X}, \boldsymbol{Y}] \\ \cdots & & \cdots \\ \underset{m\times n}{\text{cov}[\boldsymbol{Y}, \boldsymbol{X}]} & \vdots & \underset{m\times m}{\text{cov}[\boldsymbol{Y}, \boldsymbol{Y}]} \end{bmatrix}. \tag{13b}$$

Then the following properties hold.

Property 1: If $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{Z}$ where $\boldsymbol{A}$ is any nonrandom $q \times r$ matrix then, from (12),

$$\begin{aligned} \varphi_W(\boldsymbol{v}) &= E\{\exp(j\boldsymbol{v}'\boldsymbol{W})\} = E\{\exp[j(\boldsymbol{v}'\boldsymbol{A})\boldsymbol{Z}]\} \\ &= \varphi_Z(\boldsymbol{A}'\boldsymbol{v}) = \exp[j(\boldsymbol{v}'\boldsymbol{A})\boldsymbol{m} - \tfrac{1}{2}(\boldsymbol{v}'\boldsymbol{A})\boldsymbol{\Sigma}(\boldsymbol{A}'\boldsymbol{v})] \\ &= \exp[j\boldsymbol{v}'(\boldsymbol{A}\boldsymbol{m}) - \tfrac{1}{2}\boldsymbol{v}'(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')\boldsymbol{v}] \end{aligned}$$

so that $\boldsymbol{W}$ is $N(\boldsymbol{A}\boldsymbol{m}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$.

Property 2: In particular, taking $\boldsymbol{A} = [\boldsymbol{I}_n, \boldsymbol{0}]$ and then $\boldsymbol{A} = [\boldsymbol{0}, \boldsymbol{I}_m]$, where $\boldsymbol{I}_n$ is the $n \times n$ unit matrix and $\boldsymbol{0}$ is a zero matrix of the appropriate dimensions, we see with the use of (13) that the marginal distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Gaussian, i.e., $\boldsymbol{X}$ and $\boldsymbol{Y}$ are, respectively, $N(\boldsymbol{m}_X, \boldsymbol{\Sigma}_{XX})$ and $N(\boldsymbol{m}_Y, \boldsymbol{\Sigma}_{YY})$.

Property 3: Furthermore, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same dimension, taking $\boldsymbol{A} = [\boldsymbol{I}_n, \boldsymbol{I}_n]$ in Property 1 and using (13) we see that $\boldsymbol{X} + \boldsymbol{Y}$ is $N(\boldsymbol{m}_X + \boldsymbol{m}_Y, \boldsymbol{\Sigma}_{XX} + \boldsymbol{\Sigma}_{XY} + \boldsymbol{\Sigma}_{YX} + \boldsymbol{\Sigma}_{YY})$.

Property 4: If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are uncorrelated, so that $\boldsymbol{\Sigma}_{XY} = (\boldsymbol{\Sigma}_{YX})' = \boldsymbol{0}$, then they are also independent, because in this case

$$(z - \boldsymbol{m})'\boldsymbol{\Sigma}^{-1}(z - \boldsymbol{m}) = (x - \boldsymbol{m}_X)'\boldsymbol{\Sigma}_{XX}^{-1}(x - \boldsymbol{m}_X) + (y - \boldsymbol{m}_Y)'\boldsymbol{\Sigma}_{YY}^{-1}(y - \boldsymbol{m}_Y)$$

$$(2\pi)^{-r/2}|\boldsymbol{\Sigma}|^{-1/2} = (2\pi)^{-n/2}|\boldsymbol{\Sigma}_{XX}|^{-1/2}(2\pi)^{-m/2}|\boldsymbol{\Sigma}_{YY}|^{-1/2}$$

and (11) reduces to $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.

Property 5: Assuming for the moment that $\boldsymbol{m}_X = \boldsymbol{0}$ and $\boldsymbol{m}_Y = \boldsymbol{0}$, the conditional density of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is, from (11) and (13),

$$\begin{aligned} f_{X|Y}(x|y) &= f_{X,Y}(x, y)[f_Y(y)]^{-1} \\ &= K \exp -\tfrac{1}{2}[z'\boldsymbol{\Sigma}^{-1}z - y'\boldsymbol{\Sigma}_{YY}^{-1}y] \\ &= K \exp -\tfrac{1}{2}[x'\boldsymbol{S}_{XX}x + x'\boldsymbol{S}_{XY}y + y'\boldsymbol{S}_{YX}x \\ &\qquad + y'(\boldsymbol{S}_{YY} - \boldsymbol{\Sigma}_{YY}^{-1})y] \end{aligned} \tag{14}$$

where $K = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}_{YY}|^{1/2}$ and

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{XX} & \vdots & \boldsymbol{S}_{XY} \\ \boldsymbol{S}_{YX} & \vdots & \boldsymbol{S}_{YY} \end{bmatrix} = \boldsymbol{\Sigma}^{-1}$$

so that, expanding $\boldsymbol{S}\boldsymbol{\Sigma} = \boldsymbol{I}$, we have

$$\boldsymbol{S}_{XX}\boldsymbol{\Sigma}_{XX} + \boldsymbol{S}_{XY}\boldsymbol{\Sigma}_{YX} = \boldsymbol{I} \tag{15a}$$

$$S_{XX}\Sigma_{XY} + S_{XY}\Sigma_{YY} = 0 \tag{15b}$$

$$S_{YX}\Sigma_{XX} + S_{YY}\Sigma_{YX} = 0 \tag{15c}$$

$$S_{YX}\Sigma_{XY} + S_{YY}\Sigma_{YY} = I. \tag{15d}$$

Completing the square in the exponent of (14) yields

$$f_{X|Y}(x|y) = K \exp -\tfrac{1}{2}[\|x + S_{XX}^{-1}S_{XY}y\|_{S_{XX}}^2 + \|y\|^2_{S_{YY} - S_{YX}S_{XX}^{-1}S_{XY} - \Sigma_{YY}^{-1}}]. \tag{16}$$

We now note that from (15a) and (15b), respectively, we have

$$S_{XX}^{-1} = \Sigma_{XX} + S_{XX}^{-1}S_{XY}\Sigma_{YX}$$

$$S_{XX}^{-1}S_{XY} = -\Sigma_{XY}\Sigma_{YY}^{-1}$$

which may be combined to give

$$S_{XX}^{-1} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}.$$

similar
A ~~symmetric~~ argument from (15d) and (15b) yields

$$\Sigma_{YY}^{-1} = S_{YY} - S_{YX}S_{XX}^{-1}S_{XY}$$

and substitution of these in (16) shows that $f_{X|Y}(\cdot|y)$ is $N(\Sigma_{XY}\Sigma_{YY}^{-1}y,\ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX})$. More generally, if X and Y have nonzero mean, $f_{X|Y}(\cdot|y)$ is $N(m_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - m_Y),\ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX})$.

Property 6: The least squares estimator of X in terms of Y is thus the random vector

$$\hat{X} = E\{X|Y\} = m_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - m_Y). \tag{17}$$

Since this random vector is a *linear* function of the random vector Y it follows immediately from Property 1 that $E\{X|Y\}$ is $N(m_X,\ \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX})$.

Property 7: The least squares estimation error $\tilde{X} = X - E\{X|Y\}$ is the difference between two jointly Gaussian random vectors and is, therefore, from Property 3, a Gaussian random vector with mean zero (since, from Proposition 2a, the least squares estimator is unbiased) and covariance equal to the conditional covariance $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$ of X given Y (from Property 5). Thus $\tilde{X}$ is $N(0,\ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX})$.

Property 8: Any function of the least squares estimation error $\tilde{X} = X - E\{X|Y\}$ is independent of any function of the random vector Y. This follows by recalling from Proposition 2b that $\tilde{X}$ is uncorrelated with Y, and, from Property 7, that $\tilde{X}$ and Y are jointly Gaussian. Thus, from Property 4, $\tilde{X}$ and Y are independent, and invoking the standard theorem [1] that functions of independent random variables are themselves independent establishes the desired result. See [1], p. 295

From an estimation viewpoint, the most interesting and important of these properties is probably Property 6, viz., when X and Y are jointly Gaussian, the least squares estimator $\hat{X}$ of X in terms of Y is a *linear* function of Y. In fact, $\hat{X}$ is the Gaussian random vector given by (17).

If X and Y are not Gaussian random vectors, however, it is in general to be expected that the conditional expectation $E\{X|Y\}$ will be a *nonlinear* function of the random variable Y. In many such cases the calculation of the conditional mean is intractable. Under these circumstances, a common approach is to modify the problem formulation and seek the estimator $\hat{X}$ of X that is the best linear function of the random vector Y. This leads to the *linear least squares estimator* which is discussed in the next section.

V. The Linear Least Squares Estimator

In this section we restrict attention to the special case of Problem 1′ in which the estimator $\hat{X}$ is restricted to being a *linear* function of the random vector Y. The problem of interest is therefore the following.

Problem 2: Consider two jointly distributed random vectors X and Y whose means and covariances are assumed known,[2] i.e., it is assumed that the mean and covariance of the composite random vector Z defined by $Z = [X', Y']'$ are given by (13). Find the linear estimator $\hat{X} = A^\circ Y + b^\circ$ of X in terms of Y that is best in the sense that $\hat{X}$ minimizes

$$E\{\|X - AY - b\|^2\} \triangleq E\{[X - AY - b]'[X - AY - b]\} \tag{18}$$

over all linear estimators $AY + b$ of X in terms of Y, i.e., find the $n \times m$ matrix A° and the n-vector b° that minimizes (18) over all $n \times m$ matrices A and all n-vectors b. We refer to the estimator $\hat{X} = A^\circ Y + b^\circ$ that is optimal in the sense defined by this problem as the *best linear estimator* or the *linear least squares estimator* of X in terms of Y.

We remark at the outset that if its mean m_Z is known, the distribution of the random vector Z is completely specified by the distribution of its zero-mean component $Z - m_Z$, so that estimating X in terms of Y is clearly equivalent to estimating $X - m_X$ in terms of $Y - m_Y$ as long as the means m_X and m_Y are known. In other words, it can be assumed without loss of generality that X and Y have zero mean, and this we adopt as a standing assumption for the remainder of this paper, with the understanding that X and Y are to be replaced by $X - m_X$ and $Y - m_Y$ if they do not already have mean zero.

Since the trace of a real number is itself, the optimality criterion (18) is unchanged if we take the trace of both sides. Using the trace identity $\mathrm{tr}\,[FG] = \mathrm{tr}\,[GF]$ [5], [6], and the linearity of expectation and its interchangeability with the trace operation, we obtain

$$\begin{aligned} E\{\|X - AY - b\|^2\} &= \mathrm{tr}\, E\{[X - AY - b]' \cdot [X - AY - b]\} \\ &= \mathrm{tr}\, E\{[X - AY - b] \cdot [X - AY - b]'\} \\ &= \mathrm{tr}\, [E\{XX'\} - AE\{YX'\} \\ &\quad - E\{XY'\}A' + AE\{YY'\}A' \\ &\quad + bb'] \end{aligned}$$

[2] As we shall shortly see, it is not necessary to assume that the joint probability density function $f_{X,Y}(\cdot,\cdot)$ is known. It is sufficient that we know simply the means and covariances of X and Y.

$$= \text{tr}\,[\mathbf{\Sigma}_{XX} - A\mathbf{\Sigma}_{YX} - \mathbf{\Sigma}_{XY}A' + A\mathbf{\Sigma}_{YY}A' + bb'] \tag{19}$$

where for the third and fourth equalities we recall that X and Y are now assumed to have zero mean, so that $E\{bX'\} = bE\{X'\} = 0$ and $E\{XX'\} = \text{cov}\,[X, X] = \mathbf{\Sigma}_{XX}$, etc. We now note that for any positive definite matrix Q a valid definition of the inner product between two $n \times m$ matrices F and G is $\langle F, G\rangle = \text{tr}\,[FQG']$, and the corresponding induced norm is $\|F\|^2 = \text{tr}\,[FQF']$ [6]. Thus if the random vector Y is nondegenerate so that $Q = \mathbf{\Sigma}_{YY}$ is positive definite, (19) may be rewritten as

$$E\{\|X - AY - b\|^2\} = \text{tr}\,(\mathbf{\Sigma}_{XX}) - \langle A, \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\rangle - \langle \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}, A\rangle + \|A\|^2 + \|b\|^2$$

where we have used the fact that $\text{tr}\,[bb'] = \text{tr}\,[b'b] = \|b\|^2$. Completing the square then yields

$$E\{\|X - AY - b\|^2\} = \|A - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\|^2 + \|b\|^2 + \text{tr}\,[\mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}] \tag{20}$$

which is clearly uniquely minimized by taking

$$A^\circ = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}, \qquad b^\circ = \mathbf{0}.$$

Thus the least squares linear estimator is

$$\hat{X} = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}Y \tag{21}$$

and the corresponding minimum value of the variance of the estimation error $\tilde{X} = X - \hat{X}$ is, from (20),

$$E\{\|\tilde{X}\|^2\} = \text{tr}\,[\mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}].$$

In fact, it is readily found that

$$\text{cov}\,[\tilde{X}, \tilde{X}] = \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}. \tag{22}$$

In the case where X and Y do not have zero mean we have

$$\widehat{X - m_X} = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}(Y - m_Y)$$

or, equivalently

$$\hat{X} = m_X + \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}(Y - m_Y) \tag{23}$$

with the covariance of the estimation error still given by (22).

In summary, we have the following solution to Problem 2.

Proposition 3: The best linear estimator, in the sense defined by Problem 2, of the zero-mean random vector X in terms of the zero-mean random vector Y is given by

$$\hat{X} = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}Y$$

while the covariance of the corresponding estimation error

$$\tilde{X} = X - \hat{X}$$

is given by

$$\text{cov}\,[\tilde{X}, \tilde{X}] = E\{\tilde{X}\tilde{X}'\} = \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}.$$

In particular, we define the best linear *estimate* $\hat{x}$ of the value of X given that Y has value y to be

$$\hat{x} = \hat{X}(y) = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}y.$$

An alternative means of deriving this result is to set equal to zero the partial derivatives of (19) with respect to A and b, using the identities listed, for example, in [6], to obtain

$$\frac{\partial}{\partial b} E\{\|X - AY - b\|^2\} = \mathbf{0} = 2b^\circ$$

$$\frac{\partial}{\partial A} E\{\|X - AY - b\|^2\} = \mathbf{0}$$

$$= A^\circ\mathbf{\Sigma}_{YY}' + A^\circ\mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{XY} - \mathbf{\Sigma}_{YX}'$$

which may be solved to give (21). In the case where X and Y are (real-valued) random *variables* there is, of course, no need to introduce the trace operations and the process of taking partial derivatives is more direct and familiar.

It was shown in Section II that the unconstrained least squares estimator of X in terms of Y is the conditional expectation $E\{X|Y\}$, and in Section IV it was noted that when X and Y are jointly Gaussian this conditional expectation is linear in Y. Thus, when X and Y are jointly Gaussian, the unconstrained least squares estimator is already linear in Y and it must therefore coincide with the *linear* least squares estimator. A comparison of the expressions for $\hat{X}$ and $\text{cov}\,[\tilde{X}, \tilde{X}]$ given in Properties 6 and 7 of Section IV with those given above in Proposition 3 shows that this is indeed the case. In fact, the identification of the best linear estimator with the best unconstrained estimator in the Gaussian case provides a means for directly deducing several of the properties of the best linear estimator given in the following section.

In order to emphasize the fact that the best linear estimator has a number of properties, such as linearity, in common with the conditional expectation, and to provide a notation for the best linear estimator of X in terms of Y that explicitly identifies both X and Y, we introduce the notation

$$E^*\{X|Y\} = \hat{X} = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}Y. \tag{24}$$

It is emphasized that $E^*\{X|Y\}$ is simply an alternative notation for the best linear estimator of X in terms of Y: it is not to be confused with the conditional expectation $E\{X|Y\}$ with which it corresponds only in such extremely special cases as when X and Y are jointly Gaussian.

VI. Properties of the Linear Least Squares Estimator

In this section we derive a number of simple properties of the best linear estimator. As well as being important in their own right, these properties form the basis of our inductive derivation in Section VII of the discrete-time Kalman filter and predictor; in fact, this derivation is nothing more than the straightforward application of these properties to estimation problems involving linear dynamic systems. These same properties provide the basis for subsequent introduction of the so-called innovations process which plays a major role in later sections of the paper.

As before, we separate the trivial properties, which we give first, from those that are somewhat less trivial.

Proposition 4a (Properties of the Best Linear Estimator): Let X, Y, and Z be jointly distributed random vectors and let $\hat{X} = E^*\{X|Y\}$ be the best linear estimator of X in terms of Y in the sense defined by Problem 2. Then we have the following properties.

Property 1: The best linear estimator (24) and the covariance (22) of the corresponding estimation error $\tilde{X} = X - \hat{X}$ depend only on the first and second moments of the random vectors X and Y and not on their entire probability density functions. Thus jointly distributed random vectors with the same means and covariances but different probability density functions have the same estimator $\hat{X} = E^*\{X|Y\}$.

Property 2: When X and Y are jointly Gaussian the (unconstrained) least squares estimator $E\{X|Y\}$ is linear in Y and coincides with the linear least squares estimator $E^*\{X|Y\}$.

Property 3: If X and Y are uncorrelated then $E^*\{X|Y\} = E\{X\}$.

Property 4: The estimator $\hat{X} = E^*\{X|Y\}$ is unbiased in the sense that

$$E\{\tilde{X}\} = E\{X - \hat{X}\} = 0$$

or, equivalently,

$$E\{\hat{X}\} = E\{X\}.$$

Property 5: The linear least squares estimator is linear, i.e., if M is a nonrandom matrix and c is a nonrandom vector (with appropriate dimensions),

$$E^*\{MX + c|Y\} = ME^*\{X|Y\} + c$$

and the covariance of the corresponding estimation error is $M\Sigma_{\tilde{X}\tilde{X}}M'$, where $\Sigma_{\tilde{X}\tilde{X}}$ is the covariance of $X - E^*\{X|Y\}$. Also, if X and Z have the same dimension, then

$$E^*\{X + Z|Y\} = E^*\{X|Y\} + E^*\{Z|Y\}$$

and the corresponding estimation error has covariance

$$\Sigma_{(\tilde{X}+\tilde{Z})(\tilde{X}+\tilde{Z})} = \Sigma_{\tilde{X}\tilde{X}} + \Sigma_{\tilde{X}\tilde{Z}} + \Sigma_{\tilde{Z}\tilde{X}} + \Sigma_{\tilde{Z}\tilde{Z}}.$$

Property 6: For any nonnegative definite matrix F, $\hat{X} = E^*\{X|Y\}$ minimizes the weighted squared estimation error $E\{[X - g(Y)]'F[X - g(Y)]\}$ over all *linear* functions $g(Y) = AY + b$ of Y.

Proof: Properties 1 and 2 are trivial observations included for completeness. Properties 3 and 4 follow immediately on setting $\Sigma_{XY} = \text{cov}\,[X, Y] = 0$ in (24) and taking the expectation of both sides of (24) [or, more generally, (23)], respectively. Properties 5 and 6 are established by direct modification of the proof of Proposition 3.

Proposition 4b (Further Properties of the Best Linear Estimator):

Property 1: The linear least squares estimator $\hat{X} = E^*\{X|Y\}$ is characterized by the condition that the estimation error $\tilde{X} = X - \hat{X}$ is uncorrelated with Y, i.e., cov $[X - \hat{X}, Y] = 0$. In fact, $\tilde{X}$ is uncorrelated with any linear function of Y and, in particular, cov $[\tilde{X}, \hat{X}] = 0$.

Property 2: If Y and Z are uncorrelated then the best linear estimator of X in terms of both Y *and* Z (i.e., in terms of the composite vector $[Y', Z']'$) may be written

$$E^*\{X|Y, Z\} = E^*\{X|Y\} + E^*\{X|Z\} \tag{25}$$

and the corresponding estimation error has covariance $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} = \Sigma_{\tilde{X}\tilde{X}} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, where $\Sigma_{\tilde{X}\tilde{X}}$ is the covariance of $\tilde{X}_{|Y} \triangleq X - E^*\{X|Y\}$. Alternatively, with convenience for later applications in mind, we can write

$$E^*\{X|Y, Z\} = E^*\{X|Y\} + E^*\{\tilde{X}_{|Y}|Z\} \tag{26}$$

and the covariance of the estimation error as $\Sigma_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}Z}\Sigma_{ZZ}^{-1}\Sigma_{Z\tilde{X}}$ where $\Sigma_{\tilde{X}\tilde{X}}$ is as above and $\Sigma_{\tilde{X}Z} = \text{cov}\,[\tilde{X}_{|Y}, Z]$.

Property 3: If Y and Z are correlated, then the best linear estimator of X in terms of both Y and Z may be written

$$E^*\{X|Y, Z\} = E^*\{X|Y, \tilde{Z}_{|Y}\} \overset{\text{by (25)}}{=} E^*\{X|Y\} + E^*\{X|\tilde{Z}_{|Y}\}$$
$$\underset{\text{by (26)}}{=} E^*\{X|Y\} + E^*\{\tilde{X}_{|Y}|\tilde{Z}_{|Y}\}$$

Y & $\tilde{Z}_{|Y}$ are uncorrelated

where

$$\tilde{Z}_{|Y} = Z - E^*\{Z|Y\} \qquad \tilde{X}_{|Y} = X - E^*\{X|Y\}.$$

The covariance of the estimation error is given by

$$\text{cov}\,[\tilde{X}_{|Y}, \tilde{X}_{|Y}] - \text{cov}\,[\tilde{X}_{|Y}, \tilde{Z}_{|Y}][\text{cov}\,[\tilde{Z}_{|Y}, \tilde{Z}_{|Y}]]^{-1} \cdot \text{cov}\,[\tilde{Z}_{|Y}, \tilde{X}_{|Y}].$$

Property 4: More generally, the best linear estimator $\hat{X}_{|k+1} \triangleq E^*\{X|Y_1, Y_2, \cdots, Y_k, Y_{k+1}\}$ of X in terms of the random vectors $Y_1, Y_2, \cdots, Y_k, Y_{k+1}$ may be written recursively as

$$\hat{X}_{|k+1} = \hat{X}_{|k} + E^*\{\tilde{X}_{|k}|\tilde{Y}_{k+1|k}\} \tag{27}$$

where

$$\tilde{X}_{|k} \triangleq X - \hat{X}_{|k} \triangleq X - E^*\{X|Y_1, Y_2, \cdots, Y_k\} \tag{28a}$$

$$\tilde{Y}_{k+1|k} \triangleq Y_{k+1} - E^*\{Y_{k+1}|Y_1, Y_2, \cdots, Y_k\}. \tag{28b}$$

The covariance of the corresponding estimation error may be also written recursively as

$$\text{cov}\,[\tilde{X}_{|k+1}, \tilde{X}_{|k+1}] = \text{cov}\,[\tilde{X}_{|k}, \tilde{X}_{|k}] - \text{cov}\,[\tilde{X}_{|k}, \tilde{Y}_{k+1|k}] \cdot [\text{cov}\,[\tilde{Y}_{k+1|k}, \tilde{Y}_{k+1|k}]]^{-1} \cdot \text{cov}\,[\tilde{Y}_{k+1|k}, \tilde{X}_{|k}]. \tag{29}$$

Proof: 1) For any linear estimator AY of X in terms of Y we have

$$\text{cov}\,[X - AY, Y] = \Sigma_{XY} - A\Sigma_{YY}$$

which (assuming Y is nondegenerate so that Σ_{YY} is positive definite) is zero if and only if $A = \Sigma_{XY}\Sigma_{YY}^{-1}$, which, in turn, uniquely defines the best linear estimator (21). Thus cov $[X - AY, Y] = 0$ if and only if $AY = \hat{X}$ is the best linear estimator. Furthermore, for any matrix M, cov $[\tilde{X}, MY] = E\{\tilde{X}Y'\}M' = 0$.

2) Defining $W = [Y', Z']'$ we have

$$\Sigma_{WW} = \text{cov}\,[W, W] = \begin{bmatrix} \text{cov}\,[Y, Y] & \vdots & \text{cov}\,[Y, Z] \\ \cdots\cdots & \vdots & \cdots\cdots \\ \text{cov}\,[Z, Y] & \vdots & \text{cov}\,[Z, Z] \end{bmatrix} = \begin{bmatrix} \Sigma_{YY} & \vdots & 0 \\ \cdots\cdots & \vdots & \cdots\cdots \\ 0 & \vdots & \Sigma_{ZZ} \end{bmatrix} \tag{30}$$

$$\Sigma_{XW} = \text{cov}\,[X, W] = [\text{cov}\,[X, Y] \vdots \text{cov}\,[X, Z]] = [\Sigma_{XY} \vdots \Sigma_{XZ}]. \tag{31}$$

Then, using (21), we have

$$E^*\{X|Y, Z\} = E^*\{X|W\} = \Sigma_{XW}\Sigma_{WW}^{-1}W$$

See (27)

and substitution from (30) and (31) immediately yields (25). The expression for the covariance of the corresponding estimation error follows by substituting (30) and (31) into (22) with W replacing Y. The alternative expression (26) for $E^*\{X|Y, Z\}$ is a direct consequence of the observation that, writing $\hat{X}_{|Y}$ for $E^*\{X|Y\}$ and $\tilde{X}_{|Y}$ for $X - \hat{X}_{|Y}$, $\hat{X}_{|Z} = E^*\{X|Z\} = E^*\{\hat{X}_{|Y} + \tilde{X}_{|Y}|Z\} = E^*\{\tilde{X}_{|Y}|Z\}$ since $\hat{X}_{|Y}$ is a linear function of Y and Y is, by assumption, uncorrelated with Z, so that $E^*\{\hat{X}_{|Y}|Z\} = 0$. A similar argument shows that $\Sigma_{XZ} = \Sigma_{\tilde{X}Z}$ and establishes the alternative expression for the estimation error.

3) This is an immediate consequence of Property 2, the observation from Property 1 that the random vector $\tilde{Z}_{|Y} = Z - E^*\{Z|Y\}$ is uncorrelated with Y, and the observation that $E^*\{X|Y, Z\} = E^*\{X|Y, \tilde{Z}_{|Y}\}$ since a knowledge of Y and Z is clearly equivalent to a knowledge of Y and $\tilde{Z}_{|Y}$.

4) This is simply a restatement of Property 3 with Y replaced by $[Y_1', Y_2', \cdots, Y_k']'$ and Z replaced by Y_{k+1}. Q.E.D.

From the viewpoint of subsequent developments, the most important of these properties of the least squares linear estimator are its linearity, its characterization in terms of the requirement that the estimation error be uncorrelated with the data, and the "updating formulas" listed in Proposition 4b. In particular, we first have the important fact that if Y and Z are uncorrelated then the best linear estimator $E^*\{X|Y, Z\}$ of X in terms of both Y and Z can be obtained simply by additively combining the individual best linear estimators $E^*\{X|Y\}$ and $E^*\{X|Z\}$. If Y and Z are correlated, the same principle can be used once it is recalled that Y is uncorrelated with the error $\tilde{Z}_{|Y}$ in estimating Z in terms of Y, and that estimating X in terms of Y and $\tilde{Z}_{|Y}$ is the same as estimating X in terms of Y and Z, since a knowledge of Y and $\tilde{Z}_{|Y}$ is clearly equivalent to a knowledge of Y and Z, i.e., the combined vector $[Y', \tilde{Z}'_{|Y}]'$ is related to $[Y', Z']'$ by a linear transformation, and any linear function of $[Y', Z']'$ can be written as some linear function of $[Y', \tilde{Z}_{|Y}']'$ and vice versa. More generally, the best linear estimator $\hat{X}_{|k+1}$ of X in terms of the $k+1$ random vectors $Y_1, Y_2, \cdots, Y_{k+1}$ can be written recursively in terms of $\hat{X}_{|k}$ and the error vector $\tilde{Y}_{k+1|k}$ defined by (28b) using the "updating formula" (27). Thus, given the additional data in the form of another random vector Y_{k+1}, it is not necessary to resolve an entire new problem of estimating X in terms of the $k+1$ random vectors $Y_1, Y_2, \cdots, Y_k, Y_{k+1}$; all that one need do is simply to additively combine the previous best estimator $\hat{X}_{|k}$ with the best linear estimator of $\tilde{X}$ in terms of $\tilde{Y}_{k+1|k}$. As demonstrated in (29), the covariance of the corresponding estimation error may also be updated by an equally simple procedure.

The random vector $\tilde{Y}_{k+1|k}$ defined by (28b) is sometimes called the *innovation* in Y_{k+1} with respect to $Y_1, Y_2, \cdots, Y_k$ [7]–[12]. As shown in Proposition 4b, it is uncorrelated with the composite random vector $[Y_1', Y_2', \cdots, Y_k']'$ and therefore with $Y_1, Y_2, \cdots, Y_{k-1}$ and Y_k separately. It might therefore be viewed as the "component" of Y_{k+1} that conveys new information not already present in Y_1 through Y_k. This viewpoint can be made quite precise by viewing the least squares linear estimation problem as a minimum-norm problem in an appropriately defined inner product space R of random variables, and solving it using the projection theorem [13]. In this formulation, the least squares linear estimator becomes the orthogonal projection of (the components of) X on the subspace of R generated by (the components of) the random vectors $Y_1, Y_2, \cdots, Y_k, Y_{k+1}$, and the characterization of the best linear estimator as the one whose estimation error is uncorrelated with Y_1 through Y_{k+1} is merely a statement of the orthogonality of the projection, viz., that $X - \hat{X}_{|k+1}$ must be orthogonal to the vectors that generate the subspace. The iterative calculation of the innovations $Y_1, \tilde{Y}_{2|1}, \tilde{Y}_{3|2}, \cdots, \tilde{Y}_{k|k-1}, \tilde{Y}_{k+1|k}$ is nothing more than the application of the well-known Gram–Schmidt orthogonalization procedure to generate an orthogonal basis for the subspace, and because the innovations sequence $\{\tilde{Y}_{i|i-1}\}$ and the sequence of original vectors $\{Y_i\}$ generate the same subspace they convey equivalent information insofar as the linear estimation is concerned. The "updating formula" (27) is then simply a manifestation of the intuitive and easily proven observation that the projection of a vector on a subspace is the sum of its projections on each of the orthogonal basis vectors of that subspace.

We remark that if the subscript k in Y_k is interpreted to be a time index, the fact that the innovations $\tilde{Y}_{i|i-1}$ are mutually uncorrelated means that the innovations sequence $\{\tilde{Y}_{i|i-1}\}$ is discrete-time (wide-sense) white noise. Much will be made of this in the next sections when we examine estimation problems associated with linear dynamic systems.

VII. The Discrete-Time Kalman Filter and Predictor

In this section we apply the simple estimation principles developed in earlier sections to the solution of some estimation problems involving discrete-time linear dynamic systems. In particular, we consider the problem of estimating the state of a system in terms of noise-corrupted output measurements when there are random disturbances entering the state equation of the system and the initial state is a random vector. More specifically, we first direct attention to the following prediction problem formulated as Problem 3. In contrast to our earlier practice of distinguishing between random vectors and their sample values by denoting the former with capital letters and the latter with lower case letters, we henceforth follow standard practice of writing both random vectors and their sample values as lower case letters. No confusion should arise if, unless specifically indicated to the contrary, lower case letters are interpreted as random vectors.

For the first part of this section we depart temporarily from our earlier standing assumption that all vectors have zero mean, and include in the description of the dynamic system a deterministic control input and a nonzero-mean initial state. Retaining these terms in the inductive proof is no more difficult than omitting them, and it is perhaps constructive to clearly exhibit the role they play by including them in the proof.

Problem 3: Consider the discrete-time n-dimensional linear dynamic system whose state x_k at time k is generated by the difference equation

$$x_{k+1} = A_k x_k + B_k u_k + D_k \xi_k \tag{32}$$

with m-dimensional output

$$z_k = C_k x_k + \theta_k. \tag{33}$$

The initial state x_0 of the system is assumed to be a random vector with

$$E\{x_0\} = m_0 \qquad \operatorname{cov}[x_0, x_0] = \Sigma_0. \tag{34}$$

The control input $u_k \in R^r$ is assumed known (nonrandom) for all k, while the disturbances $\{\xi_k\}$ and $\{\theta_k\}$ are assumed to be uncorrelated white zero-mean random sequences with known covariances, i.e., the q- and m-dimensional random vectors ξ_k and θ_j have the following second-order statistical properties for all $k, j = 0, 1, 2, \cdots$:

$$\begin{aligned} &E\{\xi_k\} = 0 \qquad E\{\theta_k\} = 0 \\ &\operatorname{cov}[\xi_k, \xi_j] = \Xi_k \delta_{kj}, \quad \Xi_k \text{ nonnegative definite} \\ &\operatorname{cov}[\theta_k, \theta_j] = \Theta_k \delta_{kj}, \quad \Theta_k \text{ positive definite} \\ &\operatorname{cov}[\xi_k, \theta_j] = 0 \end{aligned} \tag{35}$$

where δ_{kj} is the Kronecker delta defined by

$$\delta_{kj} = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}.$$

Furthermore, it is assumed that the initial state x_0 is uncorrelated with the disturbances ξ_k and θ_k for all $k = 0, 1, 2, \cdots$, i.e.,

$$\operatorname{cov}[\xi_k, x_0] = 0 \qquad \operatorname{cov}[\theta_k, x_0] = 0. \tag{36}$$

The matrices A_k, B_k, C_k, and D_k are assumed to be known and nonrandom for all k and to have the appropriate dimensions. For each $k = 0, 1, 2, \cdots$, denote by Z_k the sequence $\{z_j\}_{j=0}{}^k$ of (random) output vectors up to and including time k, i.e.,

$$Z_k \triangleq \{z_j\}_{j=0}{}^k. \tag{37}$$

Find, for each $k = 0, 1, 2, \cdots$, the best linear estimator of the system state at time $k+1$ in terms of the output sequence up to and including time k, i.e., find the best linear estimator $\hat{x}_{k+1|k} \triangleq E^*\{x_{k+1}|Z_k\}$ of the random vector x_{k+1} in terms of the sequence Z_k of random vectors $z_0, z_1, \cdots, z_k$.

We choose to estimate the state x_{k+1} at time $k+1$ (rather than the state x_k at time k) in terms of the output sequence Z_k up to time k for convenience of later interpretations and extensions. Estimation of x_j in terms of Z_k for different relationships between j and k will be discussed later in the paper.

The solution to Problem 3 is given by the following proposition.

Proposition 5: The best linear estimator $\hat{x}_{k+1|k}$ of x_{k+1} in terms of the output sequence $Z_k = \{z_j\}_{j=0}{}^{j=k}$ may be expressed recursively using $\hat{x}_{k|k-1}$ and z_k according to the iterative relation

$$\hat{x}_{k+1|k} = A_k \hat{x}_{k|k-1} + B_k u_k + L_k [z_k - C_k \hat{x}_{k|k-1}] \tag{38}$$

with initial condition

$$\hat{x}_{0|-1} = m_0 \triangleq E\{x_0\} \tag{39}$$

where the gain matrix L_k is given for all $k = 0, 1, \cdots$ by

$$L_k = A_k \Sigma_k C_k' [C_k \Sigma_k C_k' + \Theta_k]^{-1} \tag{40}$$

and where, in turn, the $n \times n$ nonnegative definite matrix Σ_k is the covariance of the estimation error

$$\tilde{x}_{k|k-1} = x_k - \hat{x}_{k|k-1}, \tag{41}$$

i.e.,

$$\Sigma_k = \operatorname{cov}[\tilde{x}_{k|k-1}, \tilde{x}_{k|k-1}], \tag{42}$$

and may be precomputed using the iterative relation

$$\Sigma_{k+1} = A_k [\Sigma_k - \Sigma_k C_k' [C_k \Sigma_k C_k' + \Theta_k]^{-1} C_k \Sigma_k] A_k' + D_k \Xi_k D_k' \tag{43}$$

with initial condition

$$\Sigma_0 \triangleq \operatorname{cov}[x_0, x_0]. \tag{44}$$

Before proceeding with the proof of this proposition, we pause to examine the structure of the estimator it defines and, at the same time, outline the essence of an inductive proof. It should be emphasized at the outset that this estimator, which is called the Kalman (one-step) predictor [14], recursively generates $\hat{x}_{k+1|k}$ from $\hat{x}_{k|k-1}$ and the newly available measurement z_k. Its structure is shown in Fig. 1, from which we see that is it comprised of three "elements": 1) a model of the deterministic counterpart of the system (32); 2) a time-varying gain matrix L_k; and 3) a unity-gain negative feedback loop. At each time $k = 0, 1, 2, \cdots$, the unity-gain negative feedback loop generates the term $\tilde{z}_{k|k-1} = z_k - C_k \hat{x}_{k|k-1}$ which we will shortly see is uncorrelated with the past measurement sequence Z_{k-1} and is the innovation in the newly available output z_k. This is operated on by the gain L_k to generate the best linear

$Z_k = \{ z_1, \cdots, z_k \}$

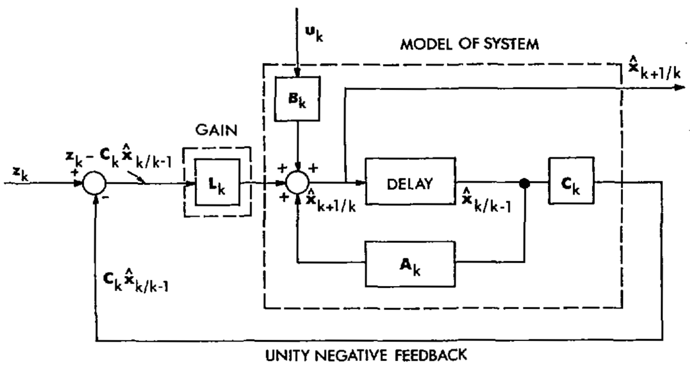


Fig. 1. The discrete-time Kalman predictor.

estimate $E^*\{x_{k+1}|\tilde{z}_{k|k-1}\} = L_k(z_k - C_k\hat{x}_{k|k-1})$ of the system state x_{k+1} in terms of the innovation $\tilde{z}_{k|k-1}$. Meanwhile, the model of the system generates the best linear estimate $\hat{x}_{k+1|k-1} = E^*\{x_{k+1}|Z_{k-1}\} = A_k\hat{x}_{k|k-1} + B_k u_k$ which, by the "updating formula" (27), may be simply added to $E^*\{x_{k+1}|\tilde{z}_{k|k-1}\}$ to give $\hat{x}_{k+1|k} = E^*\{x_{k+1}|Z_k\} = E^*\{x_{k+1}| Z_{k-1}, \tilde{z}_{k|k-1}\}$, because the innovation $\tilde{z}_{k|k-1}$ is uncorrelated with past data Z_k.

It should also be noted that the covariance Σ_k of the estimation error $\tilde{x}_{k|k-1}$ is independent of the particular control input sequence $\{u_k\}$ and may be either precomputed or computed in real time using the iterative relation (43) with boundary condition (44). Consequently the time-varying gain matrix L_k, which is derived from Σ_k via (40), is independent of the input sequence $\{u_k\}$ and may be computed in advance. Notice also that Σ_k and therefore L_k are time-varying even if the system (32) is constant and both noise sequences are stationary.

Thus the Kalman predictor that solves Problem 3 is a time-varying linear dynamic system whose dimension is that of the original system (32). It accepts as inputs the output (33) of the system (32) and the deterministic input u_k of that system, and has as its state the best linear estimate of the state of the system (32) at time $k + 1$ in terms of the particular sample output sequence Z_k received.

In the preceding section it was noted that the innovation $\tilde{z}_{k|k-1} \triangleq z_k - E^*\{z_k|Z_{k-1}\}$ is uncorrelated with all past data Z_{k-1}, and consequently the recursively generated vectors $\tilde{z}_{0|-1}(\triangleq z_0)$, $\tilde{z}_{1|0}$, $\tilde{z}_{2|1}$ $\cdots$ are mutually uncorrelated and the innovations sequence $\{\tilde{z}_{k|k-1}\}$ is (wide-sense) white noise that is related to the sequence $\{z_k\}$ by a causally invertible linear transformation. Using the linearity of the least squares linear estimator we have from (33) that

$$\hat{z}_{k|k-1} = E^*\{z_k|Z_{k-1}\} = C_k E^*\{x_k|Z_{k-1}\} + 0$$
$$= C_k\hat{x}_{k|k-1} \quad (45)$$

and we note that $E^*\{\theta_k|Z_{k-1}\}$ vanishes because Z_{k-1} depends only (and linearly) on x_0, $\{\xi_j\}_{j=0}^{k-2}$ and $\{\theta_j\}_{j=0}^{k-1}$ all of which are, by assumption, uncorrelated with θ_k. Using (45) and (33) we thus have

$$\tilde{z}_{k|k-1} \triangleq z_k - \hat{z}_{k|k-1} = z_k - C_k\hat{x}_{k|k-1} = C_k\tilde{x}_{k|k-1} + \theta_k. \quad (46)$$

Noting that θ_k is uncorrelated with both x_k and $\hat{x}_{k|k-1}$ and therefore with $\tilde{x}_{k|k-1} = x_k - \hat{x}_{k|k-1}$, a direct calculation using (46) shows that cov $[\tilde{z}_{k|k-1}, \tilde{z}_{k|k-1}] = C_k\Sigma_k C_k' + \Theta_k$, where $\Sigma_k = \text{cov}\,[\tilde{x}_{k|k-1}, \tilde{x}_{k|k-1}]$. Thus in this problem the zero-mean white-noise innovations sequence $\{\tilde{z}_{k|k-1}\}$ has covariance

$$\text{cov}\,[\tilde{z}_{k|k-1}, \tilde{z}_{j|j-1}] = [C_k\Sigma_k C_k' + \Theta_k]\delta_{kj}. \quad (47)$$

Note also that $\tilde{z}_{k|k-1}$ is uncorrelated with the future noises ξ_j for $j \geq k$ and θ_j for $j > k$. For later purposes it is also convenient to have computed that, for $k > j$,

$$\begin{aligned}
\text{cov}\,[x_k, \tilde{z}_{j|j-1}] &= \text{cov}\,[x_k, C_j\tilde{x}_{j|j-1} + \theta_j] \\
&= \text{cov}\,[x_k, \tilde{x}_{j|j-1}]C_j' + 0 \\
&= \text{cov}\,[\Phi_{k,j}x_j + \sum_{i=j}^{k-1}\Phi_{k,i+1} \\
&\qquad \cdot(B_i u_i + D_i\xi_i), \tilde{x}_{j|j-1}]C_j' \\
&= \Phi_{k,j}\,\text{cov}\,[x_j, \tilde{x}_{j|j-1}]C_j' + 0 \\
&= \Phi_{k,j}\,\text{cov}\,[\tilde{x}_{j|j-1} + \hat{x}_{j|j-1}, \tilde{x}_{j|j-1}]C_j' \\
&= \Phi_{k,j}\Sigma_j C_j' + 0 \quad (48)
\end{aligned}$$

where

$$\Phi_{k,j} = A_{k-1}A_{k-2}\cdots A_j \qquad \Phi_{j,j} = I. \quad (49)$$

Similarly, for $k > j$,

$$\text{cov}\,[\tilde{x}_{k|j-1}, \tilde{z}_{j|j-1}] = \text{cov}\,[x_k - \hat{x}_{k|j-1}, \tilde{z}_{j|j-1}]$$
$$= \Phi_{k,j}\Sigma_j C_j'. \quad (50)$$

We now turn to the proof of Proposition 5. We give two proofs, an inductive one based on the outline given in the above discussion of the structure of the Kalman predictor, and another that exploits both the properties of the innovations sequence and the characterization of the linear least squares estimator as the linear estimator whose estimation error is uncorrelated with all past data.

Proof of Proposition 5: For the inductive proof, we assume that we know $\hat{x}_{k|k-1} \triangleq E^*\{x_k|Z_{k-1}\}$ and the covariance $\Sigma_k = \text{cov}\,[\tilde{x}_{k|k-1}, \tilde{x}_{k|k-1}]$ of the corresponding estimation error $\tilde{x}_{k|k-1} \triangleq x_k - \hat{x}_{k|k-1}$. We must show that using this information and z_k we can generate $\hat{x}_{k+1|k}$ and the covariance Σ_{k+1} of the corresponding estimation error $\tilde{x}_{k+1|k}$. The inductive proof is then completed by showing that, by suitable choice of the initial conditions, the proposition gives the correct estimator for $k = 0$.

Using the "updating formula" (27), we can immediately write

$$\begin{aligned}
\hat{x}_{k+1|k} &\triangleq E^*\{x_{k+1}|Z_k\} = E^*\{x_{k+1}|Z_{k-1}, z_k\} \\
&= E^*\{x_{k+1}|Z_{k-1}\} + E^*\{\tilde{x}_{k+1|k-1}|\tilde{z}_{k|k-1}\} \\
&= \hat{x}_{k+1|k-1} + L_k[z_k - \hat{z}_{k|k-1}] \quad (51)
\end{aligned}$$

where $\tilde{z}_{k|k-1}$ is defined by (46),

$$\tilde{x}_{k+1|k-1} \triangleq x_{k+1} - \hat{x}_{k+1|k-1} \triangleq x_{k+1} - E^*\{x_{k+1}|Z_{k-1}\} \quad (52)$$

and

$$L_k = \text{cov}\,[\tilde{x}_{k+1|k-1}, \tilde{z}_{k|k-1}][\text{cov}\,[\tilde{z}_{k|k-1}, \tilde{z}_{k|k-1}]]^{-1}$$

which, using (47) and (50), becomes (40). Furthermore, from (29), the covariance of $\tilde{x}_{k+1|k} = x_{k+1} - \hat{x}_{k+1|k}$ can be written

$$\begin{aligned}\boldsymbol{\Sigma}_{k+1} \triangleq \text{cov}\,[\tilde{x}_{k+1|k}, \tilde{x}_{k+1|k}] &= \text{cov}\,[\tilde{x}_{k+1|k-1}, \tilde{x}_{k+1|k-1}]\\ &\quad - \text{cov}\,[\tilde{x}_{k+1|k-1}, \tilde{z}_{k|k-1}][\text{cov}\,[\tilde{z}_{k|k-1}, \tilde{z}_{k|k-1}]]^{-1}\\ &\quad \cdot \text{cov}\,[\tilde{z}_{k|k-1}, \tilde{x}_{k+1|k-1}]\end{aligned}$$

and, using (47) and (50), we have

$$\begin{aligned}\boldsymbol{\Sigma}_{k+1} = \text{cov}\,[\tilde{x}_{k+1|k-1}, \tilde{x}_{k+1|k-1}] - A_k\boldsymbol{\Sigma}_kC_k'\\ \cdot[C_k\boldsymbol{\Sigma}_kC_k' + \boldsymbol{\Theta}_k]^{-1}C_k\boldsymbol{\Sigma}_kA_k'. \quad (53)\end{aligned}$$

The required expressions for $\hat{x}_{k+1|k-1}$ and cov $[\tilde{x}_{k+1|k-1}, \tilde{x}_{k+1|k-1}]$ in terms of $\hat{x}_{k|k-1}$, $\boldsymbol{\Sigma}_k$ and the problem data follow trivially from the linearity of the linear least squares estimator. From Property 5 in Proposition 4a and (32) we have

$$\begin{aligned}\hat{x}_{k+1|k-1} &= A_kE^*\{x_k|Z_{k-1}\} + B_ku_k + D_kE^*\{\xi_k|Z_{k-1}\}\\ &= A_k\hat{x}_{k|k-1} + B_ku_k \quad (54)\end{aligned}$$

where $E^*\{\xi_k|Z_{k-1}\}$ vanishes because Z_{k-1} depends only (and linearly) on x_0, $\{\xi_j\}_{j=0}{}^{k-2}$ and $\{\theta_j\}_{j=0}{}^{k-1}$, all of which are, by assumption, uncorrelated with ξ_k. Furthermore, a similar argument shows that $\tilde{x}_{k|k-1}$ and ξ_k are uncorrelated, and since u_k is deterministic, the covarince of the corresponding estimation error is easily found to be

$$\begin{aligned}\text{cov}\,[\tilde{x}_{k+1|k-1}, \tilde{x}_{k+1|k-1}] &= A_k\,\text{cov}\,[\tilde{x}_{k|k-1}, \tilde{x}_{k|k-1}]A_k'\\ &\quad + D_k\,\text{cov}\,[\xi_k, \xi_k]D_k'\\ &= A_k\boldsymbol{\Sigma}_kA_k' + D_k\boldsymbol{\Xi}_kD_k' \quad (55)\end{aligned}$$

using either the formula in Property 5 of Proposition 4a or by direct calculation after subtracting (54) from (32). Substitution of (54) and (55) into (51) and (53) then yields the iterative expressions (38) and (43).

The inductive proof is completed by noting that, by choice of the boundary conditions (39) and (44), Proposition 5 holds trivially for $k = 0$, since, in the absence of any output data in terms of which to estimate x_0, $\hat{x}_{0|-1} = E\{x_0\}$ is the correct best linear estimator of x_0 and $\boldsymbol{\Sigma}_0 =$ cov $[x_0, x_0]$ the covariance of the corresponding estimation error. Q.E.D.

For the alternative proof of Proposition 5 we return to our earlier standing assumption that all random vectors have zero mean and take $u_k \equiv 0$ and $m_0 = 0$.

Alternative Proof of Proposition 5: In view of the equivalence of information (insofar as linear estimation is concerned) conveyed by the innovations sequence $\{\tilde{z}_{j|j-1}\}_{j=0}{}^k$ and the output process $\{z_j\}_{j=0}{}^k$, $\hat{x}_{k|k-1}$ may be written

$$\hat{x}_{k|k-1} = \sum_{i=0}^{k-1} G_{k,i}\tilde{z}_{i|i-1}. \quad (56)$$

The gain matrices $G_{k,i}$ can be determined using the characterization of $\hat{x}_{k|k-1}$ as the linear estimator whose estimation error is uncorrelated with all past data, i.e.,

$$\text{cov}\,[x_k - \hat{x}_{k|k-1}, \tilde{z}_{j|j-1}] = 0, \qquad \forall j \le k-1.$$

Substitution from (56) yields

$$\text{cov}\,[x_k, \tilde{z}_{j|j-1}] = \sum_{i=0}^{k-1} G_{k,i}\,\text{cov}\,[\tilde{z}_{i|i-1}, \tilde{z}_{j|j-1}], \qquad \forall j \le k-1$$

and, because $\{\tilde{z}_{j|j-1}\}$ is a white-noise process, the summation on the right reduces to a single term: in fact from (47) and (48) we have

$$G_{k,j} = \boldsymbol{\Phi}_{k,j}\boldsymbol{\Sigma}_jC_j'[C_j\boldsymbol{\Sigma}_jC_j' + \boldsymbol{\theta}_j]^{-1}. \quad (57)$$

Now note, from (49), that $\boldsymbol{\Phi}_{k+1,j} = A_k\boldsymbol{\Phi}_{k,j}$ for $j \le k-1$, while $\boldsymbol{\Phi}_{k+1,k} = A_k$, so that $G_{k+1,j} = A_kG_{k,j}$ for $j \le k-1$ and

$$\begin{aligned}\hat{x}_{k+1|k} = \sum_{j=0}^{k} G_{k+1,j}\tilde{z}_{j|j-1} &= A_k\sum_{j=0}^{k-1} G_{k,j}\tilde{z}_{j|j-1} + G_{k+1,k}\tilde{z}_{k|k-1}\\ &= A_k\hat{x}_{k|k-1} + A_k\boldsymbol{\Sigma}_kC_k'[C_k\boldsymbol{\Sigma}_kC_k' + \boldsymbol{\theta}_k]^{-1}\\ &\quad \cdot[z_k - C_k\hat{x}_{k|k-1}]\end{aligned}$$

which is (38) with $u_k = 0$. The initial condition follows as before. Subtraction of (38) from (32) yields

$$\tilde{x}_{k+1|k} = (A_k - L_kC_k)\tilde{x}_{k|k-1} - L_k\boldsymbol{\theta}_k$$

from which the difference equation (43) and boundary condition (44) for the error covariance follow by direct calculation. Thus the alternative proof is complete. Q.E.D.

Remark 1—The Gaussian Case: When the initial state x_0 and the random processes are jointly Gaussian, the system state x_{k+1} and the output sequence Z_k (or the innovations sequence $\{\tilde{z}_{j|j-1}\}_{j=0}{}^k$) are jointly Gaussian for all k, since linear transformations and sums of jointly Gaussian vectors are themselves jointly Gaussian. As noted earlier, under these circumstances the unconstrained least squares estimator is linear and thus coincides with the best linear estimator: thus the Kalman predictor defined by Proposition 5 is in this case the unconstrained least squares estimator $E\{x_{k+1}|Z_k\}$ of x_{k+1} in terms of Z_k.

Remark 2—Prediction Beyond Time $k + 1$: For any integer $i \ge 1$, the best linear estimator $\hat{x}_{k+i|k} = E^*\{x_{k+i}|Z_k\}$ of x_{k+i} in terms of Z_k may be obtained from $\hat{x}_{k+1|k}$ using the formula

$$\hat{x}_{k+i|k} = \boldsymbol{\Phi}_{k+i,k+1}\hat{x}_{k+1|k} + \sum_{j=k+1}^{k+i-1} \boldsymbol{\Phi}_{k+i,j+1}B_ju_j$$

where $\boldsymbol{\Phi}_{k,j}$ is defined by (49). This result follows by applying the linearity of the least linear estimator to the solution of (32) at time $k + i$ and recalling that, as noted earlier, Z_{k-1} is uncorrelated with ξ_j for $j \ge k$.

Remark 3—Filtering: If A_k is invertible, then $\hat{x}_{k|k} \triangleq E^*\{x_k|Z_k\}$ may be obtained from $\hat{x}_{k+1|k}$ using the relation

$$\hat{x}_{k|k} = A_k^{-1}[\hat{x}_{k+1|k} - B_k u_k].$$

More generally, whether A_k is invertible or not, $\hat{x}_{k|k}$ may be obtained directly from the linear dynamic system

$$\hat{x}_{k+1|k+1} = A_k\hat{x}_{k|k} + B_k u_k + M_{k+1}[z_{k+1} - C_{k+1}(A_k\hat{x}_{k|k} + B_k u_k)]$$

with boundary condition [recall (34)]

$$\hat{x}_{0|0} = m_0 + M_0[z_0 - C_0 m_0]$$

where

$$M_k = \Sigma_k C_k'[C_k M_k C_k' + \Theta_k]^{-1}$$

and Σ_k satisfies (43) with boundary condition (44). The proof of this assertion follows from a simple modification of the proof of Proposition 5.

Remark 4—Smoothing: Even if the matrix $\Phi_{k,j} \triangleq A_{k-1}\cdots A_j$ is invertible for all $j < k$, the best linear estimator (smoother) $\hat{x}_{j|k} \triangleq E^*\{x_j|Z_k\}$ is *not* in general given by

$$\hat{x}_{j|k} = \Phi_{k+1,j}^{-1}\left[\hat{x}_{k+1|k} - \sum_{i=j}^{k} \Phi_{k+1,i+1}B_i u_i\right] \qquad \text{(not true)}$$

if j is strictly less than k. Briefly, this is because ξ_i is clearly correlated with z_{i+1} and all future outputs $z_{i+2}, z_{i+3}, \cdots$; thus $E^*\{\xi_i|Z_k\}$ is not in general the zero vector for $i < k$. The discrete-time smoothing problem can be attacked by methods analogous to those discussed later in connection with the continuous-time state estimation problem.

Remark 5—Correlated Noises: If the noise sequences $\{\xi_k\}$ and $\{\theta_k\}$ are white, zero-mean, and uncorrelated with x_0 but correlated with each other, so that

$$\text{cov}\,[\xi_k, \theta_j] = \Gamma_k \delta_{kj},$$

(40) for L_k must be replaced by

$$L_k = [A_k\Sigma_k C_k' + D_k\Gamma_k][C_k\Sigma_k C_k' + \Theta_k]^{-1}$$

while the iterative equation (43) for Σ_{k+1} must be replaced by

$$\Sigma_{k+1} = A_k\Sigma_k A_k' - [A_k\Sigma_k C_k' + D_k\Gamma_k][C_k\Sigma_k C_k' + \Theta_k]^{-1} \cdot [C_k\Sigma_k A_k' + \Gamma_k' D_k'] + D_k\Xi_k D_k'.$$

The only modification that is necessary in the inductive proof is in the prior calculation of cov $[\tilde{x}_{k+1|k-1}, \tilde{z}_{k|k-1}]$ given by (50): the correlation between ξ_k and θ_k leads to an additional term in (50), which in this case becomes

$$\text{cov}\,[\tilde{x}_{k+1|k-1}, \tilde{z}_{k|k-1}] = A_k\Sigma_k C_k' + D_k\Gamma_k$$

and this is reflected through the two equations immediately following (52) to give the above expressions for L_k and Σ_{k+1}.

For the alternative proof, (48) must be replaced by

$$\text{cov}\,[x_k, \tilde{z}_{j|j-1}] = \Phi_{k,j+1}[A_j\Sigma_j C_j' + D_j\Gamma_j]$$

to reflect the correlation between ξ_j and θ_j, and the additional term carried through the algebra beginning with (57).

VIII. Continuous-Time Filtering, Prediction, and Smoothing

In this section we consider the continuous-time counterpart of the discrete-time state-estimation problem examined in the preceding section and derive the corresponding continuous-time Kalman–Bucy filter using the continuous-time analog of the alternative (innovations) proof given in Section VII. This derivation is formal to the extent that it involves some "familiar" formal manipulations with white noise and omits the step of rigorously proving the equivalence of the output process and the innovations process insofar as linear estimation is concerned. A more precise treatment would require the theory of stochastic integrals and stochastic differential equations (see, e.g., [15]–[18]) and would include a rigorous proof that the linear transformation that generates the innovations process from the output process is causally invertible. The innovations approach to estimation and detection problems is due to Kailath and his collaborators, to whose works the reader is referred for more detailed treatment and extensions, including nonlinear problems ([7]–[12]).

Problem 4: Consider the smooth n-dimensional linear dynamic system with state equation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + D(t)\xi(t) \qquad (58)$$

and m-dimensional output

$$z(t) = C(t)x(t) + \theta(t). \qquad (59)$$

The initial state $x(t_0)$ of the system is assumed to be a random vector with

$$E\{x(t_0)\} = m_0 \qquad \text{cov}\,[x(t_0), x(t_0)] = \Sigma_0. \qquad (60)$$

The control input $u(t) \in R^r$ is assumed known and nonrandom for all t, while the disturbances $\xi(\cdot)$ and $\theta(\cdot)$ are assumed to be white zero-mean stochastic processes that are uncorrelated with each other and with x_0, and have known covariances, i.e., the q- and m-dimensional random vectors $\xi(t)$ and $\theta(t)$ have the following second-order statistical properties for all t, s:

$$E\{\xi(t)\} = 0 \qquad E\{\theta(t)\} = 0$$

$$\text{cov}\,[\xi(t), \xi(s)] = \Xi(t)\delta(t-s), \qquad \Xi(t) \text{ nonnegative definite}$$

$$\text{cov}\,[\theta(t), \theta(s)] = \Theta(t)\delta(t-s), \qquad \Theta(t) \text{ positive definite}$$

$$\text{cov}\,[\xi(t), \theta(s)] = 0 \qquad \text{cov}\,[\xi(t), x(t_0)] = 0 \qquad \text{cov}\,[\theta(t), x(t_0)] = 0. \qquad (61)$$

The matrices $A(t)$, $B(t)$, $C(t)$, and $D(t)$ are assumed to be known and nonrandom and to have the appropriate dimensions. For each $t \geqq t_0$, denote by Z_t the stochastic process $z(\cdot)$ defined by (58) and (59) up to time t, i.e.,

$$Z_t = \{(z(s), s); s \in [t_0, t)\}. \tag{62}$$

Find, for each $t \geqq t_0$, the best linear estimator $\hat{x}(t|t) \triangleq E^*\{x(t)|Z_t\}$ of the state $x(t)$ at time t in terms of the output process Z_t up to time t.

As in the proof via the innovations approach in the preceding section, we assume without loss of generality that $u(t) \equiv 0$ and $m_0 = 0$, so that all random variables and random processes have zero mean: for convenience of reference, these terms will be reintroduced when the solution is summarized as a proposition later in the section.

Before proceeding with a derivation of the solution to Problem 4, we pause to consider what we mean by the least squares linear estimator of a random vector in terms of a continuous-time random process and to establish two subsequently needed properties that are the natural extensions to this situation of two properties of the least squares linear estimator of one random vector in terms of another. By the least squares linear estimator $E^*\{x|Y_T\}$ of an n-dimensional random vector x in terms of an m-dimensional random process $Y_T \triangleq \{(y(s), s); s \in [0, T)\}$ we mean the linear function

$$E^*\{x|Y_T\} = \hat{x} = \int_0^T H^0(\tau)y(\tau)\, d\tau \tag{63}$$

of Y_T that minimizes $E\{\|x - \int_0^T H(\tau)y(\tau)\, d\tau\|^2\}$ over all $n \times m$-matrix-valued functions $H(\cdot)$. The natural extension to this situation of the characterization of $\hat{x}$ as the linear estimator whose estimation error is uncorrelated with any linear function of the data (Proposition 4a) is that, for all $\sigma \in [0, T)$,

$$\operatorname{cov}[x - \hat{x}, y(\sigma)] = \operatorname{cov}[x, y(\sigma)] - \int_0^T H^0(\tau) \operatorname{cov}[y(\tau), y(\sigma)]\, d\tau = 0. \tag{64}$$

We establish this by observing that for any linear estimator we can write

$$\begin{aligned} E&\left\{\left\|x - \int_0^T H(\tau)y(\tau)\, d\tau\right\|^2\right\} \\ &= E\left\{\left\|x - \int_0^T H^0(\tau)y(\tau)\, d\tau - \int_0^T \tilde{H}(\tau)y(\tau)\, d\tau\right\|^2\right\} \\ &= E\left\{\left\|x - \int_0^T H^0(\tau)y(\tau)\, d\tau\right\|^2 + \left\|\int_0^T \tilde{H}(\tau)y(\tau)\, d\tau\right\|^2\right\} \\ &\quad - 2E\left\{\left[x - \int_0^T H^0(\tau)y(\tau)\, d\tau\right]' \int_0^T \tilde{H}(\sigma)y(\sigma)\, d\sigma\right\} \end{aligned} \tag{65}$$

where $H^0(\cdot)$ is defined implicitly by (64) and $\tilde{H}(t) \triangleq H(t) - H^0(t)$. Now note that the second term on the right side of (65) vanishes because it can be written

$$\operatorname{tr}\left[\int_0^T \tilde{H}(\sigma) \operatorname{cov}\left[y(\sigma), x - \int_0^T H^0(\tau)y(\tau)\, d\tau\right] d\sigma\right]$$

the integrand of which is, by (64), identically zero. The first term on the right side of (65) is independent of $H(\cdot)$, while the third term is nonnegative and can be made zero [thus minimizing (65)] by taking $\tilde{H}(\tau) \equiv 0$, i.e., $H(\tau) \equiv H^0(\tau)$. We leave it to the reader to check that if $y(\cdot)$ contains a nondegenerate white-noise component then this is the only condition under which the third term is zero, and, furthermore, under these conditions the solution to (64) is unique; thus, in this case, the least squares linear estimator $\hat{x}$ is unique and is characterized by (64). Because $\Theta(t)$ is positive definite, these conditions are met for our problem and we will shortly see, in fact, that (64) defines a unique function $H^0(\cdot)$. We note that if (64) holds for all σ then it is trivial that any linear function of Y_T is uncorrelated with $x - \hat{x}$. Note also that the linearity of the best linear estimator, viz.,

$$E^*\{Ax|Y_T\} = AE^*\{x|Y_T\} \tag{66}$$

follows immediately by observing that if $H^0(\cdot)$ satisfies (64) then $AH^0(\cdot)$ satisfies the same equation with Ax replacing x.

Returning to Problem 4, we see that since the estimation error $\tilde{z}(t|t) = z(t) - \hat{z}(t|t)$ in $\hat{z}(t|t) = E^*\{z(t)|Z_t\}$ is uncorrelated with any linear function of the past data Z_t, it is uncorrelated, in particular, with $z(\tau)$ and $\hat{z}(\tau|\tau)$ for all $\tau < t$. Thus, for all $\tau < t$, $\operatorname{cov}[\tilde{z}(t|t), \tilde{z}(\tau|\tau)] = \operatorname{cov}[\tilde{z}(t|t), z(\tau)] - \operatorname{cov}[\tilde{z}(t|t), \hat{z}(\tau|\tau)] = 0$. Clearly, by symmetry of autocorrelation functions, the same is true for $\tau > t$, so that $\operatorname{cov}[\tilde{z}(t|t), \tilde{z}(\tau|\tau)] = 0$ if $t \neq \tau$.

From the linearity of the best linear estimator and (59) we have

$$\hat{z}(t|t) = E^*\{z(t)|Z_t\} = C(t)E^*\{x(t)|Z_t\} + E^*\{\theta(t)|Z_t\} = C(t)\hat{x}(t|t) \tag{67}$$

and we note that $E^*\{\theta(t)|Z_t\}$ vanishes because Z_t depends only (and linearly) on x_0, $\{\xi(\tau); t_0 \leq \tau < t\}$ and $\{\theta(\tau); t_0 \leq \tau < t\}$, all of which are, by assumption, uncorrelated with $\theta(t)$. Thus the innovation $\tilde{z}(t|t)$ may be written

$$\begin{aligned} \tilde{z}(t|t) &= z(t) - C(t)\hat{x}(t|t) = C(t)[x(t) - \hat{x}(t|t)] + \theta(t) \\ &= C(t)\tilde{x}(t|t) + \theta(t) \end{aligned} \tag{68}$$

from which it can be calculated directly that $\operatorname{cov}[\tilde{z}(t|t), \tilde{z}(t|t)]$ is an impulse with magnitude $\Theta(t)$, using the fact that $\theta(t)$ is uncorrelated with both $\hat{x}(t|t)$ (because, as discussed above, it is uncorrelated with Z_t) and $x(t)$, and therefore with $\tilde{x}(t|t)$. Thus the zero-mean white-noise innovations process $\tilde{z}$ has covariance

$$\operatorname{cov}[\tilde{z}(t|t), \tilde{z}(\tau|\tau)] = \Theta(t)\delta(t - \tau). \tag{69}$$

As in Section VII, it is convenient for later purposes to have calculated by augments similar to those leading to (48) and (50) that

$$\operatorname{cov}[x(t), \tilde{z}(\tau|\tau)] = \Phi(t, \tau)\Sigma(\tau)C'(t) = \operatorname{cov}[\tilde{x}(t|\tau), \tilde{z}(\tau|\tau)] \tag{70}$$

where $\Sigma(t) = \operatorname{cov}[\tilde{x}(t|t), \tilde{x}(t|t)]$ and $\Phi(t, \tau)$ is the transition matrix associated with the differential equation $\dot{x}(t) = A(t)x(t)$.

Assuming that the linear transformation that generates the innovations process $\tilde{z}$ from the output process z is causally invertible, so that any linear function of z can be expressed as a linear function of $\tilde{z}$ (and vice versa), we can write

$$\hat{x}(t|t) = \int_{t_0}^{t} G(t,\tau)\tilde{z}(\tau|\tau)\,d\tau \tag{71}$$

where the matrix-valued gain function $G(t,\cdot)$ can be calculated using the characterization (64) of $\hat{x}(t|t)$, viz.,

$$\operatorname{cov}[x(t), \tilde{z}(\sigma|\sigma)] = \int_{t_0}^{t} G(t,\tau)\operatorname{cov}[\tilde{z}(\tau|\tau), \tilde{z}(\sigma|\sigma)]\,d\tau, \qquad t_0 \le \sigma < t.$$

Using (69) and (70) this reduces to

$$G(t,\sigma) = \Phi(t,\sigma)\Sigma(\sigma)C'(\sigma)\Theta^{-1}(\sigma), \qquad t_0 \le \sigma < t. \tag{72}$$

Formally differentiating (71) with respect to t using the Leibnitz rule, (72), and the defining property of the transition matrix, viz, $(d/dt)\Phi(t,\tau) = A(t)\Phi(t,\tau)$, we obtain

$$\begin{aligned}\dot{\hat{x}}(t|t) &= A(t)\int_{t_0}^{t} G(t,\tau)\tilde{z}(\tau|\tau)\,d\tau + G(t,t)\tilde{z}(t|t)\\ &= A(t)\hat{x}(t|t) + G(t,t)\tilde{z}(t|t)\end{aligned} \tag{73}$$

with, setting $t = t_0$ in (71), initial condition $\hat{x}(t_0|t_0) = \mathbf{0}$. Note that (72) defines $G(t,\sigma)$ only for $\sigma < t$ and yet (73) requires that we determine $G(t,t)$. If we integrate (73) from t_0 to t we obtain

$$\hat{x}(t|t) = \int_{t_0}^{t} \Phi(t,\tau)G(\tau,\tau)\tilde{z}(\tau|\tau)\,d\tau$$

and if this expression is to coincide with (71) when $G(t,\tau)$ is defined for $\tau < t$ by (72), we must have

$$\Phi(t,\tau)G(\tau,\tau) = \Phi(t,\tau)\Sigma(\tau)C'(\tau)\Theta^{-1}(\tau)$$

which, since $\Phi(t,\tau)$ is nonsingular, yields

$$G(\tau,\tau) = \Sigma(\tau)C'(\tau)\Theta^{-1}(\tau) \tag{74}$$

and (73) becomes

$$\dot{\hat{x}}(t|t) = A(t)\hat{x}(t|t) + \Sigma(t)C'(t)\Theta^{-1}(t)\,[z(t) - C(t)\hat{x}(t|t)]. \tag{75}$$

Subtracting this equation from (58) (with $u(t) \equiv 0$) gives

$$\dot{\tilde{x}}(t|t) = [A(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)]\tilde{x}(t|t) + D(t)\xi(t) - \Sigma(t)C'(t)\Theta^{-1}(t)\theta(t)$$

from which we can write

$$\tilde{x}(t|t) = \psi(t,t_0)\tilde{x}(t_0|t_0) + \int_{t_0}^{t} \psi(t,\tau)[D(\tau)\xi(\tau) - \Sigma(\tau)C'(\tau)\Theta^{-1}(\tau)\theta(\tau)]\,d\tau \tag{76}$$

where $\psi(t,\tau)$ is the transition matrix associated with $\dot{w}(t) = [A(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)]w(t)$. A direct calculation using the fact that θ and ξ are white and uncorrelated with each other and $\tilde{x}(t_0|t_0)(= x(t_0))$ then shows that

$$\begin{aligned}\Sigma(t) \triangleq \operatorname{cov}[\tilde{x}(t|t), \tilde{x}(t|t)] &= \psi(t,t_0)\Sigma_0\psi'(t,t_0)\\ &+ \int_{t_0}^{t}\psi(t,\tau)\Sigma(\tau)C'(\tau)\Theta^{-1}(\tau)C(\tau)\Sigma(\tau)\psi'(t,\tau)\,dt\\ &+ \int_{t_0}^{t}\Psi(t,\tau)D(\tau)\Xi(\tau)D'(\tau)\Psi'(t,\tau)\,dt\end{aligned} \tag{77}$$

and this may be differentiated to give

$$\begin{aligned}\dot{\Sigma}(t) &= [A(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)]\Sigma(t)\\ &\quad + \Sigma(t)[A(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)]'\\ &\quad + D(t)\Xi(t)D'(t) + \Sigma(t)C'(t)\Theta^{-1}(t)C(t)\Sigma(t)\\ &= A(t)\Sigma(t) + \Sigma(t)A'(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)\Sigma(t)\\ &\quad + D(t)\Xi(t)D'(t)\end{aligned} \tag{78}$$

and, setting $t = t_0$ in (77), the boundary condition $\Sigma(t_0) = \Sigma_0$.

A nonzero control input and a nonzero-mean initial state simply affect the mean of $x(t)$, which is then

$$E\{x(t)\} = \Phi(t,t_0)m_0 + \int_{t_0}^{t}\Phi(t,\tau)B(\tau)u(\tau)\,d\tau$$

and this sum must be added to the right side of (71). This leads to an additional term $B(t)u(t)$ on the right side of (75) and changes its initial condition from zero to m_0. The differential equations for $\tilde{x}(t|t)$ and $\Sigma(t)$ remain unaffected.

In summary, we have the following solution to Problem 4.

Proposition 6: The best linear estimator $\hat{x}(t|t) = E^*\{x(t)|Z_t\}$ of the state $x(t)$ of the system (58) in terms of the output process Z_t is the n-dimensional linear dynamic system

$$\dot{\hat{x}}(t|t) = A(t)\hat{x}(t|t) + B(t)u(t) + \Sigma(t)C'(t)\Theta^{-1}(t)\cdot[z(t) - C(t)\hat{x}(t|t)]$$

with initial condition

$$\hat{x}(t_0|t_0) = m_0 = E\{x(t_0)\}$$

where the $n \times n$ symmetric nonnegative definite matrix $\Sigma(t)$ is the covariance of the estimation error $\tilde{x}(t|t) = x(t) - \hat{x}(t|t)$ and is the solution to the Riccati equation

$$\dot{\Sigma}(t) = A(t)\Sigma(t) + \Sigma(t)A'(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)\Sigma(t) + D(t)\Xi(t)D'(t)$$

with initial condition

$$\Sigma(t_0) = \Sigma_0 = \operatorname{cov}[x(t_0), x(t_0)].$$

The form of this best linear estimator, which is called the Kalman–Bucy filter for Problem 4, is shown in Fig. 2 [19]. Like its discrete-time counterpart, it is comprised of three "elements."

1) A unity-gain negative feedback loop which generates

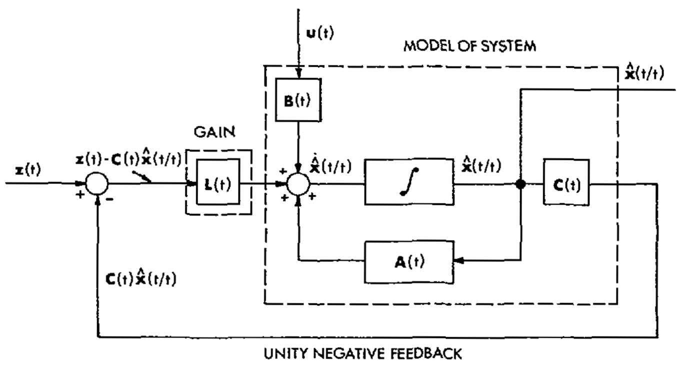


Fig. 2. The continuous-time Kalman–Bucy filter.

at each time t the innovation $\tilde{z}(t|t) = z(t) - C(t)\hat{x}(t|t)$ in the newly available output measurement $z(t)$.

2) A time-varying gain $L(t) = \Sigma(t)C'(t)\Theta^{-1}(t)$ which operates on the innovation $\tilde{z}(t|t)$ to give what can be interpreted as the best linear estimator of $\dot{x}(t)$ minus its mean in terms of $\tilde{z}(t|t)$.

3) A model of the deterministic part of the system (58), the internal feedback loop and externally applied input of which give $A(t)\hat{x}(t|t) + B(t)u(t)$, which can be interpreted as the best linear estimator of $\dot{x}(t)$ prior to the arrival of $z(t)$, and which, when additively combined with $L(t)\tilde{z}(t|t)$, gives $\dot{\hat{x}}(t|t) = (d/dt)E^*\{x(t)|Z_t\}$.

Notice also that, as in the discrete-time case, the covariance $\Sigma(t)$ of the estimation error is independent of the deterministic control input and may be either precomputed or computed in real time from the Riccati equation (78) with initial condition $\Sigma(t_0) = \Sigma_0$. The existence of solutions to this Riccati equation and the behavior of the solution as t approaches infinity will be examined in the next section.

Remark 6—The Gaussian Case: If $x(t_0)$, $\xi(\cdot)$ and $\theta(\cdot)$ are jointly Gaussian so also are the state $x(t)$ and the innovations process $\tilde{z}(\cdot|\cdot)$ [or the output process $z(\cdot)$], and the Kalman–Bucy filter is the unconstrained least squares estimator of $x(t)$ in terms of Z_t, i.e., $\hat{x}(t|t) = E\{x(t)|Z_t\}$.

Remark 7—Prediction: For $T \geqq 0$, the best linear estimator of $x(t+T)$ in terms of Z_t may be obtained from $\hat{x}(t|t)$ using the relation

$$E^*\{x(t+T)|Z_t\} \triangleq \hat{x}(t+T|t) = \Phi(t+T,t)\hat{x}(t|t) + \int_t^{t+T} \Phi(t+T,s)B(s)\,u(s)\,ds$$

which follows immediately by applying the linearity of the best linear estimator to the solution of (58) and recalling that, as noted earlier, $\xi(s)$ is uncorrelated with Z_t for $s \geq t$. A direct calculation shows that the covariance of the corresponding estimation error is given by

$$\text{cov}\,[\tilde{x}(t+T|t), \tilde{x}(t+T|t)] = \Phi(t+T,t)\Sigma(t)\Phi'(t+T,t) + \int_t^{t+T} \Phi(t+T,\tau)D(\tau)\Xi(\tau)D'(\tau)\Phi'(t+T,\tau)\,d\tau.$$

These expressions are *not* valid in the smoothing case where $T < 0$, which will be discussed in a later remark.

Remark 8—Correlated Noises: If $\xi(\cdot)$ and $\theta(\cdot)$ are correlated with each other so that

$$\text{cov}\,[\xi(t), \theta(s)] = \Gamma(t)\delta(t-s)$$

then equations (78) and (75) must be replaced by

$$\begin{aligned}\dot{\Sigma}(t) &= A(t)\Sigma(t) + \Sigma(t)A'(t) - [\Sigma(t)C'(t) \\ &\quad + D(t)\Gamma(t)]\Theta^{-1}(t)[C(t)\Sigma(t) + \Gamma'(t)D'(t)] \\ &\quad + D(t)\Xi(t)D'(t) \\ \dot{\hat{x}}(t|t) &= A(t)\hat{x}(t|t) + B(t)u(t) \\ &\quad + [\Sigma(t)C'(t) + D(t)\Gamma(t)]\Theta^{-1}(t)[z(t) - C(t)\hat{x}(t|t)].\end{aligned}$$

Otherwise, Proposition 6 remains unchanged. These modifications follow from the observation that (70) must be replaced by

$$\text{cov}\,[x(t), \tilde{z}(\tau|\tau)] = \Phi(t,\tau)[\Sigma(\tau)C'(\tau) + D(\tau)\Gamma(\tau)]$$

and this change reflected through the subsequent algebra; otherwise, the proof is unchanged.

Remark 9—Colored Noises: Consider now the case where, in addition to being correlated, the noise processes $\xi(\cdot)$ and $\theta(\cdot)$ are not white. This problem can be reduced immediately to an equivalent higher dimensional problem involving white-noise processes, the solution to which is discussed in Remark 8, provided that the combined process

$$n(\cdot) = [\xi'(\cdot), \theta'(\cdot)]' \tag{79}$$

is a finite-dimensional Markov process and $\theta(\cdot)$ contains a white component with nonsingular covariance matrix, i.e., provided that the combined process $n(\cdot)$ can be generated as the output of a finite-dimensional linear dynamic system of the form

$$\dot{q}(t) = F(t)q(t) + G(t)w(t) \tag{80a}$$

$$\begin{aligned} n(t) = \begin{bmatrix}\xi(t)\\ \theta(t)\end{bmatrix} &= \begin{bmatrix}H_1(t)\\ H_2(t)\end{bmatrix} q(t) + \begin{bmatrix}J_1(t)\\ J_2(t)\end{bmatrix} v(t) \\ &= H(t)q(t) + J(t)v(t) \end{aligned} \tag{80b}$$

where $w(\cdot)$ and $v(\cdot)$ are (possibly correlated) white-noise processes with covariances

$$\text{cov}\,[w(t), w(\tau)] = W(t)\delta(t-\tau)$$

$$\text{cov}\,[v(t), v(\tau)] = V(t)\delta(t-\tau)$$

and, in addition,

$$J_2(t)V(t)J_2'(t) > 0, \qquad \forall t. \tag{81}$$

The requirement (81) that $\theta(t)$ contain a nondegenerate white-noise component will be seen shortly to correspond to our earlier standing assumption (61) that $\Theta(t)$ be positive definite when $\theta(\cdot)$ is white.

Under these circumstances, the systems (58) and (80) can be combined into the single system

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ \cdots \\ q(t) \end{bmatrix} = \begin{bmatrix} A(t) & \vdots & D(t)H_1(t) \\ \cdots & \vdots & \cdots \\ 0 & \vdots & F(t) \end{bmatrix} \begin{bmatrix} x(t) \\ \cdots \\ q(t) \end{bmatrix} + \begin{bmatrix} B(t) \\ \cdots \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} D(t)J_1(t) & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & G(t) \end{bmatrix} \begin{bmatrix} v(t) \\ \cdots \\ w(t) \end{bmatrix} \tag{82a}$$

$$z(t) = [C(t) : H_2(t)] \begin{bmatrix} x(t) \\ \cdots \\ q(t) \end{bmatrix} + J_2(t)v(t) \tag{82b}$$

which involves white (but correlated) noises in the dynamics and the measurements. The Kalman filter and predictor for estimating the state of this combined system is then obtained as a direct application of Remark 8, and the estimate of $x(t)$ is obtained from this estimate of the state of the combined system using the linearity of the best linear estimator, i.e.,

$$\hat{x}(t|t) = E^*\{x(t)|Z_t\} = [I:0]E^*\left\{\begin{bmatrix} x(t) \\ \cdots \\ q(t) \end{bmatrix} \middle| Z_t\right\}.$$

We note that the requirement (81) ensures that the additive white noise in the measurements (82) is nondegenerate in the sense that it has positive definite covariance matrix, which conforms with our earlier standing assumption (61) on the additive white measurement noise.

If we restrict attention to white-noise processes $v(\cdot)$ and $w(\cdot)$ that are uncorrelated, a necessary and sufficient condition for the combined process $\mathfrak{n}(\cdot)$ defined by (79) to be representable as the output of a finite-dimensional linear dynamic system of the form (82) is that its autocorrelation function $R(t, \tau) = E\{\mathfrak{n}(t)\mathfrak{n}'(\tau)\}$ can be expressed for all t and τ as

$$R(t, \tau) = P(t)Q(t \wedge \tau)P'(\tau) + S(t)S'(\tau)\delta(t - \tau) \tag{83}$$

for some continuous matrix-valued functions $P(\cdot)$ and $S(\cdot)$ and some continuously differentiable, symmetric, nonnegative definite, matrix-valued function $Q(\cdot)$ whose derivative $\dot{Q}(\cdot)$ also has nonnegative definite values. The notation $t \wedge \tau$ denotes the minimum of t and τ [20]–[22].

If the combined process $\mathfrak{n}(\cdot)$ is wide-sense stationary, so that $R(t, \tau) = R(t - \tau, 0) = R'(\tau - t, 0)$, an equivalent characterization can be given in the frequency domain in terms of the Laplace transform $\hat{R}(\cdot)$ of $R(\cdot, 0)$: a necessary and sufficient condition for the wide-sense stationary process $\mathfrak{n}(\cdot)$ to be representable as the output of a stable, constant finite-dimensional linear dynamic system of the form (82) is that its spectrum $\hat{R}(\cdot) = \mathcal{L}[E\{\mathfrak{n}(t)\mathfrak{n}'(0)\}]$ is a rational function of the form

$$\hat{R}(s) = \frac{Q(s)Q'(-s)}{\rho(s)\rho(-s)} + TT' \tag{84}$$

for some polynomial $\rho(s) = s^n + \sum_{i=0}^{n-1} \alpha_i s^i$ with all roots having negative real parts, some matrix-valued polynomial $Q(s) = \sum_{i=0}^{n-1} Q_i s^i$ with degree at most $n - 1$ and all zeros in the left half-plane, and some matrix T [21]–[24].

Remark 10—Smoothing: Consider now the problem of finding $\hat{x}(t|T) = E^*\{x(t)|Z_T\}$, where $T > t$. Proceeding as in the filtering case, we first write

$$\hat{x}(t|T) = \int_{t_0}^{T} H(t, T; \tau)\tilde{z}(\tau|\tau)\, d\tau \tag{85}$$

and then apply the characterization of $\hat{x}(t|T)$ as the linear estimator whose estimation error is uncorrelated with all the data up to time T to obtain

$$\begin{aligned} \operatorname{cov}[x(t), \tilde{z}(\sigma|\sigma)] &= \int_{t_0}^{T} H(t, T; \tau) \operatorname{cov}[\tilde{z}(\tau|\tau), \tilde{z}(\sigma|\sigma)]\, d\tau \\ &= H(t, T; \sigma)\Theta(\sigma), \qquad t_0 \le \sigma < T. \end{aligned}$$

Now observe that for $\sigma < t$ this equation is identical to (72) and $H(t, T; \sigma) = G(t, \sigma) = \Phi(t, \sigma)\Sigma(\sigma)C'(\sigma)$, so that splitting the integral into two parts, we can write (85) as

$$\begin{aligned} \hat{x}(t|T) &= \int_{t_0}^{t} G(t, \tau)\tilde{z}(\tau|\tau)\, d\tau \\ &\quad + \int_{t}^{T} \operatorname{cov}[x(t), \tilde{z}(\tau|\tau)]\Theta^{-1}(\tau)\tilde{z}(\tau|\tau)\, d\tau \\ &= \hat{x}(t|t) + \int_{t}^{T} \operatorname{cov}[x(t), \tilde{z}(\tau|\tau)]\Theta^{-1}(\tau)\tilde{z}(\tau|\tau)\, d\tau. \end{aligned} \tag{86}$$

Now for $t < \tau$ we have that

$$\begin{aligned} \operatorname{cov}[x(t), \tilde{z}(\tau|\tau)] &= \operatorname{cov}[x(t), \tilde{x}(\tau|\tau)]C'(\tau) + 0 \\ &= \operatorname{cov}[\tilde{x}(t|t) + \hat{x}(t|t), \tilde{x}(\tau|\tau)]C'(\tau) \\ &= \operatorname{cov}[\tilde{x}(t|t), \tilde{x}(\tau|\tau)]C'(\tau) + 0 \\ &\triangleq P(t, \tau)C'(\tau) \end{aligned} \tag{87}$$

and, using (76), we calculate, for $t \le \tau$,

$$P(t, \tau) \triangleq \operatorname{cov}[\tilde{x}(t|t), \tilde{x}(\tau|\tau)] = \Sigma(t)\psi'(\tau, t), \qquad t \le \tau \tag{88a}$$

where, as before, $\psi(\tau, t)$ is the transition matrix associated with $\dot{w}(t) = [A(t) - \Sigma(t)C'(t)\Theta^{-1}(t)C(t)]w(t)$. We note in passing that, for $t \ge \tau$, we have

$$P(t, \tau) \triangleq \operatorname{cov}[\tilde{x}(t|t), \tilde{x}(\tau|\tau)] = \psi(t, \tau)\Sigma(\tau), \qquad t \ge \tau. \tag{88b}$$

Thus

$$\hat{x}(t|T) = \hat{x}(t|t) + \Sigma(t)\int_{t}^{T} \psi'(\tau, t)C'(\tau)\Theta^{-1}(\tau)\,[z(\tau) - \hat{z}(\tau|\tau)]\, d\tau \tag{89}$$

and subtracting each side of this equation from $x(t)$ yields

$$\tilde{x}(t|T) = \tilde{x}(t|t) - \Sigma(t)\int_{t}^{T} \psi'(\tau, t)C'(\tau)\Theta^{-1}(\tau)\tilde{z}(\tau|\tau)\, d\tau$$

from which a direct calculation using (69) and (88) shows

$$\boldsymbol{\Gamma}(t) \triangleq \operatorname{cov}\,[\tilde{x}(t|T), \tilde{x}(t|T)] = \boldsymbol{\Sigma}(t) - \boldsymbol{\Sigma}(t)\left[\int_t^T \boldsymbol{\psi}'(\tau, t)\boldsymbol{C}'(\tau)\boldsymbol{\Theta}^{-1}(\tau)\boldsymbol{C}(\tau)\boldsymbol{\psi}(\tau, t)\, d\tau\right]\boldsymbol{\Sigma}(t). \tag{90}$$

If T is fixed and we differentiate (89) with respect to t using (75), (78), and the identity [2] $d/dt\boldsymbol{\psi}'(\tau, t) = -[\boldsymbol{A}(t) - \boldsymbol{\Sigma}(t)\boldsymbol{C}'(t)\boldsymbol{\Theta}^{-1}(t)\boldsymbol{C}(t)]'\boldsymbol{\psi}'(\tau, t)$, we find after some algebra and the reuse of (89) that

$$\dot{\hat{x}}(t|T) = \boldsymbol{A}(t)\hat{x}(t|T) + \boldsymbol{D}(t)\boldsymbol{\Xi}(t)\boldsymbol{D}'(t)\boldsymbol{\Sigma}^{-1}(t) \cdot [\hat{x}(t|T) - \hat{x}(t|t)] \tag{91}$$

which is integrated backwards from the final time T, with the boundary condition $\hat{x}(T|T)$ obtained by integrating (75) forward from t_0 to t. Notice that once $\hat{x}(t|t)$ has been found over the entire interval from t_0 to T there is no need to retain the output measurements, since they are not needed in (91). We remark that $\boldsymbol{\Sigma}^{-1}(t)$ can be computed directly from

$$\dot{\boldsymbol{\Sigma}}^{-1} = -\boldsymbol{A}'\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A} - \boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Xi}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1} + \boldsymbol{C}'\boldsymbol{\Theta}^{-1}\boldsymbol{C}$$

$$\boldsymbol{\Sigma}^{-1}(t_0) = \boldsymbol{\Sigma}_0^{-1}$$

which follows from (78) and the identity [2] $d/dt(\boldsymbol{\Sigma}^{-1}(t)) = -\boldsymbol{\Sigma}^{-1}(t)\dot{\boldsymbol{\Sigma}}(t)\boldsymbol{\Sigma}^{-1}(t)$. Differentiation of (90) shows, after some algebra, that $\boldsymbol{\Gamma}(t)$ satisfies

$$\dot{\boldsymbol{\Gamma}}(t) = [\boldsymbol{A}(t) + \boldsymbol{D}(t)\boldsymbol{\Xi}(t)\boldsymbol{D}'(t)\boldsymbol{\Sigma}^{-1}(t)]\boldsymbol{\Gamma}(t) + \boldsymbol{\Gamma}(t)[\boldsymbol{A}(t) + \boldsymbol{D}(t)\boldsymbol{\Xi}(t)\boldsymbol{D}'(t)\boldsymbol{\Sigma}^{-1}(t)]' - \boldsymbol{D}(t)\boldsymbol{\Xi}(t)\boldsymbol{D}'(t)$$

with terminal condition $\boldsymbol{\Gamma}(T) = \boldsymbol{\Sigma}(T)$.

Various other representations of the solution to the smoothing problem have been proposed when the problem is specialized to the so-called fixed-interval, fixed-point, or fixed-lag smoothing problems. A summary of these can be found in [9], including the two-filter solution to the fixed-interval problem given in [27].

IX. The Duality Between Least Squares Estimation and Least Squares Control

Consider the least squares regulator problem involving the linear dynamic system

$$\dot{\boldsymbol{p}}(t) = -\boldsymbol{A}'(t)\boldsymbol{p}(t) - \boldsymbol{C}'(t)\boldsymbol{w}(t) \tag{92a}$$

$$\boldsymbol{v}(t) = \boldsymbol{D}'(t)\boldsymbol{p}(t) \tag{92b}$$

operating in reverse time from the boundary condition

$$\boldsymbol{p}(t_1) = \boldsymbol{p}_1 \tag{93}$$

and the quadratic cost functional

$$J[\boldsymbol{w}] = \int_{t_1}^{t_0} [\boldsymbol{w}'(t)\boldsymbol{\Theta}(t)\boldsymbol{w}(t) + \boldsymbol{v}'(t)\boldsymbol{\Xi}(t)\boldsymbol{v}(t)]\, d(-t) + \boldsymbol{p}'(t_0)\boldsymbol{\Sigma}_0\boldsymbol{p}(t_0)$$
$$= \int_{t_0}^{t_1} [\boldsymbol{w}'(t)\boldsymbol{\Theta}(t)\boldsymbol{w}(t) + \boldsymbol{p}'(t)\boldsymbol{D}(t)\boldsymbol{\Xi}(t)\boldsymbol{D}'(t)\boldsymbol{p}(t)]\, dt + \boldsymbol{p}'(t_0)\boldsymbol{\Sigma}_0\boldsymbol{p}(t_0). \tag{94}$$

It is well known [2] that the control law $\boldsymbol{w}(\cdot, \cdot): R^n \times [t_0, t_1] \rightarrow R^m$ that minimizes $J[\boldsymbol{w}]$ for all boundary conditions $\boldsymbol{p}(t_1) = \boldsymbol{p}_1$ is given by

$$\boldsymbol{w}^0(\boldsymbol{p}(t), t) = -\boldsymbol{L}'(t)\boldsymbol{p}(t) = -\boldsymbol{\Theta}^{-1}(t)\boldsymbol{C}(t)\boldsymbol{K}(t)\boldsymbol{p}(t) \tag{95}$$

where the $n \times n$ nonnegative definite matrix $\boldsymbol{K}(t)$ satisfies the Riccati equation

$$\dot{\boldsymbol{K}}(t) = \boldsymbol{A}(t)\boldsymbol{K}(t) + \boldsymbol{K}(t)\boldsymbol{A}'(t) - \boldsymbol{K}(t)\boldsymbol{C}'(t)\boldsymbol{\Theta}^{-1}(t)\boldsymbol{C}(t)\boldsymbol{K}(t) + \boldsymbol{D}(t)\boldsymbol{\Xi}(t)\boldsymbol{D}'(t) \tag{96a}$$

with initial condition (terminal condition in reverse time)

$$\boldsymbol{K}(t_0) = \boldsymbol{\Sigma}_0. \tag{96b}$$

Furthermore, the corresponding minimum cost is

$$J[\boldsymbol{w}^0] = \boldsymbol{p}_1'\boldsymbol{K}(t_1)\boldsymbol{p}_1. \tag{97}$$

It should be noted at the outset that the Riccati equation (96a) and boundary condition (96b) for $\boldsymbol{K}(\cdot)$ are identical to the Riccati equation and boundary condition (78) for $\boldsymbol{\Sigma}(\cdot)$ in the Kalman–Bucy filter of Proposition 6. Furthermore, making the consequent identification of $\boldsymbol{K}(t)$ with $\boldsymbol{\Sigma}(t)$, it is seen that the time-varying gain matrix $\boldsymbol{L}'(t) = \boldsymbol{\Theta}^{-1}(t)\boldsymbol{C}(t)\boldsymbol{K}(t)$ in the negative feedback loop that implements the optimum control law (95) for the least squares control problem is simply the transpose of the time-varying gain matrix $\boldsymbol{L}(t) = \boldsymbol{\Sigma}(t)\boldsymbol{C}'(t)\boldsymbol{\Theta}^{-1}(t)$ that operates on the innovation $\boldsymbol{z}(t) - \boldsymbol{C}(t)\hat{x}(t|t)$ in the Kalman–Bucy filter of Proposition 6. It can thus be seen that there is a one-to-one correspondence between the solutions of least squares estimation problems of the type discussed in Section VIII and the solutions of least squares control problems of the type discussed above. Partly because dual (or adjoint) systems are involved, this correspondence is often referred to as the duality between least squares estimation and least squares control.

In this vein, we remark that the linear dynamic system defined by (92) with boundary condition at the terminal time t_1 is the *dual* (or *adjoint*) of the linear dynamic system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}(t)\boldsymbol{x}(t) + \boldsymbol{D}(t)\boldsymbol{u}(t) \tag{98a}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}(t)\boldsymbol{x}(t) \tag{98b}$$

with boundary condition at the initial time t_0. Thus the system considered in the above least squares control problem is the dual of the system involved in the least squares estimation problem of Section VIII. Furthermore, the Kalman–Bucy filter of Proposition 6, viz.,

$$\dot{\hat{x}}(t|t) = [\boldsymbol{A}(t) - \boldsymbol{L}(t)\boldsymbol{C}(t)]\hat{x}(t|t) + \boldsymbol{L}(t)\boldsymbol{z}(t) \tag{99}$$

with boundary condition at the initial time t_0 has as its dual the system

$$(t) = -[\boldsymbol{A}(t) - \boldsymbol{L}(t)\boldsymbol{C}(t)]'\boldsymbol{p}(t) \tag{100a}$$

$$\boldsymbol{v}(t) = -\boldsymbol{L}'(t)\boldsymbol{p}(t) \tag{100b}$$

with boundary condition at the terminal time t_1. The state equation of this dual system will be immediately recognized as the optimum closed-loop system for the above least squares control problem, while the output equation

TABLE I

Least Squares Estimation Problem		Least Squares Control Problem	
System	$\dot{x} = Ax + D\xi$, $z = Cx + \theta$ $(\mathcal{L})$	System	$\dot{p} = -A'p - C'w$, $v = D'p$ $(\mathcal{L}^*)$
Boundary condition	At t_0	Boundary condition	At t_1
Covariances	$\operatorname{cov}[x(t_0), x(t_0)] = \Sigma_0$, $\operatorname{cov}[\xi(t), \xi(\tau)] = \Xi\delta(t-\tau)$, $\operatorname{cov}[\theta(t), \theta(\tau)] = \Theta\delta(t-\tau)$	Cost Functional	$J[w] = \int_{t_0}^{t_1} w'\Theta w + v'\Xi v\, dt + p'(t_0)\Sigma_0 p(t_0)$
Solution	$\dot{\hat{x}} = A\hat{x} + L(z - C\hat{x})$ $(\mathcal{S})$, $L = \Sigma C'\Theta^{-1}$, $\dot{\Sigma} = A\Sigma + \Sigma A' - \Sigma C'\Theta^{-1}C\Sigma + D\Xi D'$, $\Sigma(t_0) = \Sigma_0$	Solution	$\dot{p}^0 = -(A - LC)'p^0$ $(\mathcal{S}^*)$, $w^0 = -L'p^0$, $L = KC'\Theta^{-1}$, $\dot{K} = AK + KA' - KC'\Theta^{-1}CK + D\Xi D'$, $K(t_0) = \Sigma_0$
Transition matrix	$\phi_A(t, s) = e^{A(t-s)}$	Transition matrix	$\phi_{A'}(t, s) = e^{A'(s-t)}$
Controllability matrix	$[D, A, D, \cdots, A^{n-1}D]$ $\mathcal{L}$ completely controllable	Observability matrix	$[D, A D, \cdots, A^{n-1}D]$ $\mathcal{L}^*$ completely observable
Observability matrix	$[C', A'C', \cdots, (A')^{n-1}C']$ $\mathcal{L}$ completely observable	Controllability matrix	$[C', A'C', \cdots, (A')^{n-1}C']$ $\mathcal{L}^*$ completely controllable

defines the corresponding optimum control, i.e., $v(t) = w^0(p(t), t) = -L'(t)p(t)$. Thus the solution to the above least squares control problem is given by the dual of the solution to the least squares estimation problem of Section VIII.

The duality between estimation and control provides a direct and convenient means for examining the properties of the Kalman–Bucy filter and its associated Riccati equation by drawing on established properties of the solution to the dual least squares control problem. In particular, we can examine the existence and properties of any steady-state solution to the Kalman filtering problem by appealing to the known solution of the corresponding infinite-time regulator problem. With this in mind, we identify in Table I the corresponding properties of the two problems and their solutions. Since our greatest subsequent interest is in the case where the systems are constant and the noises are stationary or, equivalently, the weighting matrices Ξ and Θ in the cost functional are constant, we restrict attention to this situation.

The following properties of the solution to the Riccati equation (96) with initial condition Σ_0 are well known for the least squares control problem. We exploit the duality discussed above to write these properties directly in terms of the notation of the estimation problem, and include in brackets the equivalent conditions on the system involved in the control problem from which these properties are deduced. Of particular importance here is the well-known fact that a constant system is completely controllable (respectively completely observable) if and only if its dual is completely observable (respectively completely controllable); this is evident from the last two lines of Table I.

For fixed but arbitrary t, the solution $\Sigma_{t_0}(t)$ to the Riccati equation (96) with boundary condition $\Sigma_{t_0}(t_0) = 0$ has the following properties as the initial time t_0 decreases to increasingly negative values.

1) If the constant system (98a) is completely observable from the output (98b) [equivalently, if the dual system (92) is completely controllable from the input w], then $\Sigma_{t_0}(t)$ is monotone nondecreasing as t_0 decreases, and is uniformly bounded from above. As $t_0 \to -\infty$, $\Sigma_{t_0}(t)$ approaches a limit Σ_∞ that is independent of t and is a (not necessarily unique) nonnegative definite solution to the algebraic Riccati equation

$$A\Sigma_\infty + \Sigma_\infty A' - \Sigma_\infty C'\Theta^{-1}C\Sigma_\infty + D\Xi D' = 0. \quad (101)$$

2) If, in addition, the constant system (98) is completely controllable from the input u [equivalently, if the dual system (92a) is completely observable from the output (92b)], then $\Sigma_{t_0}(t)$ is positive definite for all $t > t_0$ and Σ_∞ is the unique positive definite solution to the algebraic Riccati equation (101).

3) If the constant system (98) [equivalently, the dual system (92)] is both completely controllable and completely observable, then the eigenvalues of $A - \Sigma_\infty C'\Theta^{-1}C$ have strictly negative real parts, so that the system

$$\dot{\hat{x}}(t|t) = A\hat{x}(t|t) + \Sigma_\infty C'\Theta^{-1}[z(t) - C\hat{x}(t|t)] \quad (102)$$

is asymptotically stable, and so bounded inputs result in bounded outputs, operating in forward time. [Equivalently, the dual system

$$\dot{p}(t) = -[A - \Sigma_\infty C'\Theta^{-1}C]'p(t)$$

is asymptotically stable when operating in reverse time.]

With these facts at hand, it is a simple matter to show that the constant feedback control law

$$w^0(p(t), t) = -L_\infty'p(t) = -\Theta^{-1}C\Sigma_\infty p(t) \quad (103)$$

is the unique solution to the infinite-time regulator problem defined by the constant completely controllable and completely observable system (92) and the cost functional (94), where the weighting matrices Θ and Ξ are constant and $t_0 = -\infty$ (see, e.g., [2]). The natural analog of this result in the estimation case is that the steady-state Kalman filter (102) is the best linear estimator of the state of the completely controllable and completely observable constant system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + D\xi(t) \\ z(t) &= Cx(t) + \theta(t) \end{aligned} \quad (104)$$

in terms of the output process $z(\cdot)$ over $(-\infty, t)$. Since, however, the system (104) has been operating "since $-\infty$," the processes $x(\cdot)$ and $z(\cdot)$ will not be well defined unless the stability properties of the system are such that the covariance of $x(t)$ has reached a steady-state value $\operatorname{cov}[x(t), x(t)] = \psi_x$ which is given by the solution to the algebraic matrix equation

$$A\psi_x + \psi_x A' = -D\Xi D'.$$

A sufficient condition for there to exist a unique such solution is that all the eigenvalues of A have strictly negative real parts. [2]

Under these circumstances the problem becomes identical to the classical Wiener filtering problem, and the steady-state Kalman filter (102) is the optimum realizable Wiener filter for this problem.

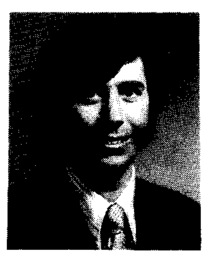
Perhaps of more practical importance, however, is that the steady-state Kalman filter (102) is trivially the solution to the finite-time estimation problem of Section VIII if the covariance at the finite initial time t_0 is taken to be cov $[x(t_0), x(t_0)] = \boldsymbol{\Sigma}_\infty$. Also of importance is the result that, even if cov $[x(t_0), x(t_0)]$ is not $\boldsymbol{\Sigma}_\infty$ but some other value $\boldsymbol{\Sigma}_0$, the error in using the steady-state Kalman filter (102) instead of the correct time-varying filter approaches zero as $t \rightarrow \infty$ [25], [26].

Acknowledgment

The comments and suggestions of D. L. Snyder, T. Kailath, and A. S. Gilman are gratefully acknowledged.

References

[1] E. Parzen, *Modern Probability Theory and Its Applications.* New York: Wiley, 1960.
[2] R. W. Brockett, *Finite Dimensional Linear Systems.* New York: Wiley, 1970.
[3] W. M. Wonham, "On the separation theorem of stochastic control," *SIAM J. Contr.*, vol. 6, no. 2, pp. 312–326, 1968.
[4] H. Cramer, *Mathematical Methods of Statistics.* Princeton, N.J.: Princeton University Press, 1946.
[5] R. Bellman, *Introduction to Matrix Analysis*, 2nd ed. New York: McGraw-Hill, 1970.
[6] M. Athans, "The matrix minimum principle," *Inform. Contr.*, vol. 11, pp. 592–606, Nov./Dec. 1967.
[7] T. Kailath, "The innovations approach to detection and estimation theory," *Proc. IEEE*, vol. 58, pp. 680–695, May 1970.
[8] ——, "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.
[9] T. Kailath and P. Frost, "An innovations approach to least-squares estimation—Part II: Linear smoothing in additive white noise," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 655–660, Dec. 1968.
[10] 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.
[11] T. Kailath and R. Geesey, "An innovations approach to least-squares estimation, Part IV: Recursive estimation given lumped covariance functions," this issue, pp. 720–727.
[12] P. Frost, "Nonlinear estimation in continuous-time systems," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, Calif., June 1968.
[13] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: Wiley, 1969.
[14] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME*, vol. 82, ser. D, pp. 35–43, Mar. 1960.
[15] K. Ito and H. P. McKean, *Diffusion Processes and Their Sample Paths.* New York: Springer, 1965.
[16] H. P. McKean, *Stochastic Integrals.* New York: Academic Press, 1969.
[17] I. I. Gikhman and A. V. Skorokhod, *Introduction to the Theory of Random Processes.* Philadelphia: Saunders, 1969.
[18] E. Wong, *Stochastic Processes in Information and Dynamic Systems.* New York: McGraw-Hill, 1971.
[19] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Trans. ASME*, vol. 83, ser. D, pp. 95–108, Mar. 1961.
[20] 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.
[21] R. Geesey, "Canonical representations of second-order processes with applications," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, Calif., Dec. 1968.
[22] T. Kailath and R. Geesey, "Covariance factorization—An explication via examples," in *Proc. 2nd Asilomar Conf. Systems and Circuits*, Monterey, Calif., Nov. 1968.
[23] D. C. Youla, "On the factorization of rational matrices," *IRE Trans. Inform. Theory*, vol. IT-7, pp. 172–189, July 1961.
[24] B. D. O. Anderson, "An algebraic solution to the spectral factorization problem," *IEEE Trans. Automat. Contr.*, vol. AC-12, pp. 410–414, Aug. 1967.
[25] B. Friedland, "On the effect of incorrect gain in Kalman filter," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-12, p. 610, Oct. 1967.
[26] R. A. Singer and P. A. Frost, "On the relative performance of the Kalman and Wiener filters," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 390–394, Aug. 1969.
[27] D. C. Fraser and J. E. Potter, "The optimum linear smoother as a combination of two optimum linear filters," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 387–390, Aug. 1969.



Ian B. Rhodes (M'67) was born in Melbourne, Australia, on May 29, 1941. He received the B.E. and M. Eng. Sc. degrees in electrical engineering from the University of Melbourne, Melbourne, Australia, in 1963 and 1965, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, Calif., in 1968.

In January 1968 he was appointed Assistant Professor of Electrical Engineering at the Massachusetts Institute of Technology and taught there until September 1970, when he joined the faculty of Washington University, St. Louis, Mo., as Associate Professor of Engineering and Applied Science in the graduate program of Control Systems Science and Engineering. His current research interests lie in linear system theory, optimal estimation and control, and minimax problems.

Dr. Rhodes is a member of the Society of Industrial and Applied Mathematics, the Institute of Radio and Electronics Engineers (Australia), and Sigma Xi, and is an Associate Editor of the IFAC journal *Automatica.*