M. Gevers P7
privé 010/656064
proj " /41.81.81-, ext. 2590

AN INNOVATIONS APPROACH TO THE DISCRETE-TIME LINEAR LEAST-SQUARES ESTIMATION PROBLEM

by

W.R.E. Wouters* and M. Gevers**

SUMMARY :

The "innovations" concept and the notion of an "innovations process" associated with a stochastic process was first introduced in 1968 and has since proven to be useful in a wide variety of fields such as least-squares estimation, detection, representation of stochastic processes, identification of processes, etc... One of the major contributions of the innovations approach, perhaps, has been the greater intuitive insight it has provided in the study of stochastic processes and in the derivation of least-squares estimation formulas.

In this paper we present in a (hopefully) tutorial and unifying way most of the important concepts and results on innovations and linear least-squares estimation which are now scattered throughout the literature.

---

* Department of Mathematics and Systems Engineering
Koninklijke/Shell-Laboratorium, Amsterdam (Shell Research B.V.)

** Laboratoire d'Automatique et d'Analyse des Systèmes, Université Catholique de Louvain, Bâtiment Maxwell, 1348 Louvain-la-Neuve (Belgium).

1. INTRODUCTION

1.1. General introduction.

The linear least-squares estimation problem in continuous time was tackled by Wiener [1] in the 1940's. His approach was to obtain the impulse response of the linear estimation filter as the solution of an integral equation involving the autocovariance function of the observed process. A solution of this integral equation was difficult to obtain except in some simple cases, among them the very important stationary scalar case where the optimal filter can be obtained via a factorization of the spectral density function of the observed process.

It was not until 1960 that the nonstationary (and later also nonlinear) least-squares estimation problems could be successfully solved. This was due to the introduction by Kalman [2] and Kalman and Bucy [3] of recursive statespace models and algorithms. The methods are restricted to a class of processes, namely those that can be modelled by a white noise driven through a "lumped" dynamical system (i.e. a system that can be described by a finite number of paramters, or elements). Such processes are actually functions (and in the linear case projections) of Markov processes. However, more important than the restriction to a class of lumped processes is the fact that the Kalman filter solution to the least-squares estimation problem requires a complete knowledge of the model, thus sidestepping the difficult problem of obtaining linear least-squares estimates from covariance information. The discrete-time Kalman filter was derived using (at least implicitly) the innovations concept, although it was not called so. However, the extension to continuous time required a tedious and physically not very intuitive limiting technique.

The innovations approach to the solution of linear least-squares estimation problems can be traced back to Kolmogorov in 1941 [4]. Kolmogorov studied only discrete-time problems, and he solved them using using a orthonormalization procedure to "whiten" the observations. With whitened observations, the resulting linear least-squares estimation problems are much simpler to solve.

Motivated at first by the desire to obtain the Kalman-Bucy filter using Kolmogorov's approach, Kailath and his associates [5] - [12] introduced and popularized the "innovations" concept in the field of least-squares estimation and detection, thereby providing a unifying and physically intuitive means of solving a wide range of problems, for both continuous and discrete-time processes, stationary and nonstationary, Gaussian and non-Gaussian.

The innovations approach could also have been called the "whitening" approach. The basic idea is that, for a large class of stochastic processes, one can associate with a given stochastic process $\{y\}$ a related white noise process $\{\varepsilon\}$ that can be obtained from $\{y\}$ by a causal and causally invertible transformation. All that is needed to make this transformation is the knowledge of the covariance function of the process $\{y\}$. By a causal and causally invertible transformation we mean that given a particular realization $\{y_t\}$ of the process $\{y\}$ over the interval $\{0,\tau\}$ (in continuous or discrete time), one can compute the associated white noise realization $\{\varepsilon_t\}$, $t \in [0,\tau]$, by passing the observations y_t through a causal filter that is a function of the covariance function of the process $\{y\}$. Conversely, given the white noise realization $\{\varepsilon_t\}_{t=0}^{\tau}$, one can derive the observation record $\{y_t\}_{t=0}^{\tau}$ by passing the "innovations" $\{\varepsilon_t\}_{t=0}^{\tau}$ through a causal filter which is the inverse of the first one (hence the term causally invertible). This filter (or system) driven by the white noise innovations can therefore be used as a model for the process $\{y\}$. Hence the term <u>innovations representation</u> of $\{y\}$.

In the innovations approach, then, the observed process is first converted to its associated white noise innovations process, which can then be considered as a new observation process. The idea behind this conversion is that estimation and detection problems are much easier to solve when the observed process is white; the innovation ε_t can be looked upon as the "new information" about y_t, i.e. it is that part of y_t that cannot be predicted from the past observation record.

In addition to providing a physical interpretation of and an elegant means of deriving most estimation and detection formulas, the innovations approach has enabled the derivation of a number of new and sometimes simpler expressions which otherwise would have been very difficult to arrive at. Some of these will be given in this paper. The method has also been applied to the solution of several nonlinear least-squares estimation problems [8] and to large classes of Gaussian and non-Gaussian detection problems [13]. Finally, the innovations approach to estimation has notably helped researchers rediscover that covariance information is all that is needed for linear least-squares estimation. After the attempts to solve the Wiener-Hopf equation had been overshadowed by the enormous success of the Kalman filter formulation, the idea permeated most engineering circles that a lumped model of the observation process was necessary to derive recursive estimation formulas for a related signal process. This is not true of course, since the Wiener-Hopf equation (even if it cannot often be solved) and elementary estimation theory tell us that linear least-squares estimation filters depend only upon the covariance functions. The innovations approach has thus contributed to bridge this gap, since it has provided a solution to the problem of deriving a lumped white-noise-driven model from the covariance function [9],[11]. This problem is called the stochastic realization problem, and will be treated in a separate paper.

1.2. Outline of this paper.

The purpose of this paper and its companion [14] is to try to present in two papers and in a (hopefully) tutorial way most of the important concepts and results on innovations, which are now scattered in a large number of publications. To keep things simple, we shall in this paper be dealing only with linear and discrete-time problems. We should say at the outset that the results and derivations we present here are well-known and most of our material is based on [5], [6], [11], [12]. It is our hope that the reader will want to learn more about innovations by reading some of the references.

We begin in Section 2 by defining the innovations process associated with a finite-variance observation process $\{y_t\}_{t=0}^{\infty}$. It is shown that $\{\varepsilon_t\}_{t=0}^{\tau}$ is an uncorrelated (white) sequence which contains the same statistical information as the sequence $\{y_t\}_{t=0}^{\tau}$, in the sense that these two sequences can be obtained from one another through causal and causally invertible filters. The expression for these filters is computed in terms of the covariance function of {y}. Next we show how linear least-squares estimates of any random process {x} can be obtained from the innovations process {ε} associated with the observation process {y}. These estimates are obtained as the output of what we call the general innovations filter (GIF), whose inputs are the innovations. The filter is expressed entirely in terms of the autocovariance of {y} and the cross covariance between {x} and {y}. In Section 3, the estimation formulas obtained for the GIF are specialized for the case where the process {y} is the output of a known state-space model. For such a case we obtain the finitely recursive Kalman-Bucy formulas for least-squares filtering and prediction of the state of the model. The more complicated smoothing formulas are subsequently derived, once again as a special application of the GIF formulas. To conclude this section, we show how one can very easily obtain the estimate of any process that is linearly related to the state x. Section 4 again deals with a special use of the GIF formulas, this time for the case where the process {y} is the output of a known autoregressive moving average (ARMA) model. We show how finitely recursive formulas can be obtained for the predicted values of y.

Throughout this paper we have tried to stress the generality of the GIF; all the lumped model formulas are derived as applications of the more general GIF formulas.

2. THE GENERAL INNOVATIONS FILTER IN DISCRETE TIME.

2.1. Linear Least-Squares Estimation - The Projection Theorem.

Before we give a definition of innovations, or derive any filter formulas, we do want to state and prove the projection theorem, which forms the basis of all linear least-squares (LLS) estimation theory. In LLS estimation we consider problems of the following nature : given a set of finite variance, p-vector random variables $\{y_s\}_{s=0}^{t}$, find a LLS estimate of a n-vector random variable, say x_τ.

By the linearity assumption, we require our estimator to have the form

$$\hat{x}_{\tau/t} = \sum_{s=0}^{t} A_s\, y_s \qquad (2.1)$$

which can be written as

$$\hat{x}_{\tau/t} = A\, Y^t \qquad (2.2)$$

where

$$Y^t \triangleq \left[y_t^T,\ y_{t-1}^T \cdots,\ y_o^T\right]^T \qquad (2.3)$$

a p(t+1) vector of random variables and $A \triangleq [A_t \quad A_{t-1} \cdots \quad A_o]$ a n x p(t+1) matrix of constants.
The estimation error will be defined by

$$\tilde{x}_{\tau/t} = x_\tau - \hat{x}_{\tau/t} = x_\tau - A\, Y^t \qquad (2.4)$$

The Projection Theorem can then be formulated as follows :

Projection Theorem

$\hat{x}_{\tau/t}$ as defined by (2.2) is a least-squares estimate of x_τ if and only if the estimation error $\tilde{x}_{\tau/t}$ as given by (2.4) is orthogonal (uncorrelated) to the random variables $\{y_s\}_{s=0}^{t}$, on which the estimation is based, i.e.

$$E\{\tilde{x}_{\tau/t}\, y_s^T\} = 0, \qquad 0 \leq s \leq t$$

Proof : Let us consider the optimization problem :

$$\min_A \; J(A) \triangleq E\{(x_\tau - A\, Y^t)^T (x_\tau - A\, Y^t)\}$$

From this definition it is clear that J(A) is a scalar convex quadratic function of A. A necessary and sufficient condition for the optimum is that $\frac{\partial J}{\partial A} \equiv 0$. Using the standard formula for the derivative of a scalar function with respect to a matrix we obtain

$$\frac{\partial J(A)}{\partial A} = -2\, E\{\tilde{x}_{\tau/t} (Y^t)^T\} = 0$$

From the definition of Y^t the theorem follows. A very important corollary of the projection theorem will be formulated presently.

Corollary.

If $\hat{x}_{\tau/t}$ is the LLS of x_τ based upon the random variables $\{y_s\}_{s=0}^t$, then it follows that the estimation error $\tilde{x}_{\tau/t}$ is orthogonal (uncorrelated) to any linear combination of the set $\{y_s\}_{s=0}^t$.

Proof : The proof is obvious, and will be omitted. The theorem and its corollary are very important however, and will be used again and again through this paper. A very good outline of the geometrical interpretation of the projection theorem in connection with LLS estimation is given in [15 , chapter 4]. Another aspect of the projection theorem is that it has a very intuitive appeal, and is relatively well known to engineers. Indeed, the fact that the estimation error should be uncorrelated with the data on which the estimation is based is intuitively logical, since one can argue that if any such correlation were left, part of the error could have been foreseen, so the estimate can't be the "best".

2.2. The one step ahead prediction problem - Definition of the innovations.

In the earlier part of this section we introduced a set of random variables $\{y_t\}$. From here on, we shall think of $\{y_t\}$ as a stochastic process. Let us assume then that we have a zero mean observation process of second order $\{y_t\}_{t=0}^{\infty}$, and that its autocovariance function $\{R_y(t,\tau),\ \tau,t \geq 0\}$ is known. We will address the problem of finding a linear least-squares estimate of y_t, given the observations $y_o, y_1, \ldots, y_{t-1}$. This LLS estimate will be denoted $\hat{y}_{t/t-1}$. Doing so for t=0,1,2,... etc., we obtain a sequence of related variables, namely the error in the estimates, i.e.

$$\varepsilon_t = y_t - \hat{y}_{t/t-1} \quad ; \quad t \geq 0 \ ; \ \hat{y}_{0/-1} = 0 \qquad (2.5)$$

The stochastic process defined in this way is called the innovations process. It is fairly straightforward to show that the innovations process as defined by equation (2.5) is an orthogonal process, i.e. a white noise. Indeed, if we consider the correlation $E\{\varepsilon_t \varepsilon_\tau^T\}$, $\tau<t$ say, then we know from the corollary that $\varepsilon_t = y_t - \hat{y}_{t/t-1}$ is uncorrelated with any linear combination of $\{y_s\}_{s=0}^{t-1}$. Now, $\varepsilon_\tau = y_\tau - \hat{y}_{\tau/\tau-1}$ $(\tau<t)$ where $\hat{y}_{\tau/\tau-1}$ is a linear combination of $\{y_s\}_{s=0}^{\tau-1} \subseteq \{y_s\}_{s=0}^{t-1}$ and so it follows that ε_τ is a linear combination of $\{y_s\}_{s=0}^{\tau} \subseteq \{y_s\}_{s=0}^{t-1}$, whence $E\{\varepsilon_t \varepsilon_\tau^T\} = 0$. Of course, the argument is also valid when we consider $E\{\varepsilon_t \varepsilon_\tau^T\}$ with $t<\tau$, since in that case we have that ε_τ must be orthogonal to all past observations, and that ε_t is a linear combination of the same.

Now that we have defined the innovations process, let us try to derive a linear, innovations driven, filter to generate the sequence of estimates $\hat{y}_{t/t-1}$.

We are thus looking for coefficients $C_{t,\tau}$ (actually these coefficients can be interpreted as impulse response elements) such that

$$\hat{y}_{t/t-1} = \sum_{\tau=0}^{t-1} C_{t,\tau}\, \varepsilon_\tau \qquad (2.6)$$

is a linear least-squares estimate of y_t. We know that the optimality of $\hat{y}_{t/t-1}$ must imply that $\varepsilon_t = y_t - \hat{y}_{t/t-1}$ is uncorrelated with all previous ε_τ, thus

$$E\{(y_t - \hat{y}_{t/t-1})\, \varepsilon_s^T\} = 0 \; ; \quad 0 < s < t \qquad (2.7)$$

which yields the equality

$$R_{y\varepsilon}(t,\tau) = C_{t,\tau}\, R_\varepsilon(\tau,\tau) \qquad (2.8)^\dagger$$

In this paper and in part II [14] we will always demand that the stochastic process $\{y_t\}$ is of full rank, which implies that the matrix $R_\varepsilon(\tau,\tau)$ is non-singular for all $\tau \geq 0$. The coefficients $C_{t,\tau}$ may then be obtained from (2.8) by postmultiplying the equation with the inverse matrix $R_\varepsilon^{-1}(\tau,\tau)$:

$$C_{t,\tau} = R_{y\varepsilon}(t,\tau)\, R_\varepsilon^{-1}(\tau,\tau) \qquad (2.9)$$

Our desired expression for the innovations driven predictor then becomes :

$$\hat{y}_{t/t-1} = \sum_{\tau=0}^{t-1} R_{y\varepsilon}(t,\tau)\, R_\varepsilon^{-1}(\tau,\tau)\, \varepsilon_\tau \; ; \; t \geq 0 \; ; \; \hat{y}_{0/-1} = 0 \qquad (2.10)$$

The required "input" sequence, $\{\varepsilon_\tau\}$ for the filter is determined by equation (2.5) which defines the innovations.

The predictor arrangement is depicted in figure 2.1.

† The capital letter R is used to denote covariance functions. The presence of two subscripts indicates a cross covariance, i.e. $R_{y\varepsilon}(t,\tau) \triangleq E\{y_t \varepsilon_\tau^T\}$. A single subscript is used to refer to the autocovariance function, i.e. $R_y(t,\tau) \triangleq E\{y_t y_\tau^T\}$.

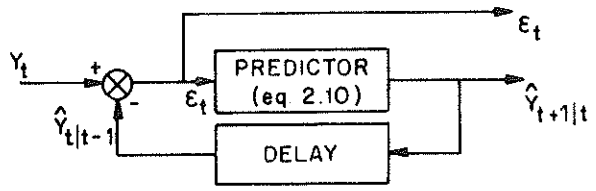

Fig. 2.1. Innovations driven predictor for y_t .

The equations developed so far show that we can indeed transform the original observation process $\{y_t\}$ by a causal filter into a "whitened" version, the innovations process $\{\varepsilon_t\}$. We will presently show that this transformation is also causally invertible. To see this, we have to show that from the record of whitened observations $\{\varepsilon_t\}$ and the knowledge of the predictor equation (2.10), we can actually reconstruct the observations sequence $\{y_t\}$. That this is indeed so, can be most easily demonstrated by looking at figure 2.2, and reinterpreting equation (2.5) to read

$$y_t = \hat{y}_{t/t-1} + \varepsilon_t \qquad t \geq 0, \quad \hat{y}_{0/-1} = 0 \qquad (2.5')$$

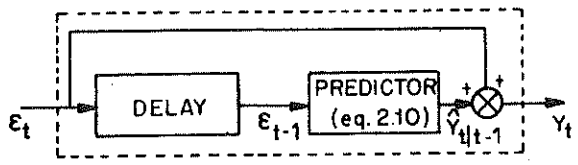

Fig. 2.2. The inverse filter generating $\{y_t\}$ from $\{\varepsilon_t\}$.

The causal system depicted inside the dotted line in Figure 2.2. may be described by the equation (2.11) and is also referred to as the innovations representation (IR) of the process $\{y_t\}$. We have thus proved the important fact that every finite variance discrete-time stochastic process admits an innovations representation.

$$y_t = \sum_{\tau=0}^{t} R_{y\varepsilon}(t,\tau)\, R_{\varepsilon}^{-1}(\tau,\tau)\varepsilon_\tau \qquad (2.11)$$

The only thing left to do now, is to derive suitable expressions for the covariance function $R_{y\varepsilon}(t,\tau)$ and the innovations variance function $R_{\varepsilon}(\tau,\tau)$, as a function of the process covariance function $R_y(t,\tau)$. As it is, we will show that $R_{y\varepsilon}(t,\tau)$ and $R_{\varepsilon}(\tau,\tau)$ can be computed recursively from $R_y(t,\tau)$.

A recursion for $R_{y\varepsilon}(t,\tau)$ follows immediately from

$$R_{y\varepsilon}(t,\tau) = E\left\{y_t \left(y_\tau - \sum_{s=0}^{\tau-1} R_{y\varepsilon}(\tau,s)\, R_{\varepsilon}^{-1}(s,s)\, \varepsilon_s\right)^T\right\}$$

$$R_{y\varepsilon}(t,\tau) = R_y(t,\tau) - \sum_{s=0}^{\tau-1} R_{y\varepsilon}(t,s)\, R_{\varepsilon}^{-1}(s,s)\, R_{y\varepsilon}^{T}(\tau,s) \qquad (2.12)$$

$$t,\tau > 0$$

whereby the variance of the innovations process is given by

$$R_{\varepsilon}(t,t) = R_{y\varepsilon}(t,t) \qquad (2.13)$$

This last equality follows from

$E\{\hat{y}_{t/t-1}\, \varepsilon_t^T\} = 0$, by virtue of the fact that $\hat{y}_{t/t-1}$ is a linear combination of $\{\varepsilon_s\}_{s=0}^{t-1}$, and the fact that $\{\varepsilon_t\}$ is an orthogonal process.

This leaves us with $E\{\varepsilon_t \varepsilon_t^T\} = E\{(y_t - \hat{y}_{t/t-1})\varepsilon_t^T\} = R_{y\varepsilon}(t,t)$.

Formulas (2.10), (2.12) and (2.13) completely define the predicted estimates $\{y_{t/t-1}\}$ as a function of the innovations, and the autocovariance function of the observations process. Furthermore one sees that the formulas (2.10), (2.12) and (2.13) constitute a recursive solution to the problem. In general, however, this solution is not finitely recursive, and so a growing memory is required. One case where we do obtain finite recursions is when the autocovariance function $R_y(t,\tau)$ has the "truncation" property, i.e. $R_y(t,\tau) = 0$ for all $|\tau-t| > M$; $M < \infty$.

2.3. Estimation of a related process - The general innovations filter (GIF).

So far we have defined the innovations process and we have also shown that any discrete time process $\{y_t\}$ admits an innovations representation. Next we have developed recursive formulas for the predicted estimate. Now we will consider the estimation of a related process $\{x_t\}$, again based on the observations process $\{y_t\}$. The cross covariance function $\{R_{xy}(t,\tau)$; $t,\tau > 0\}$ is supposed to be known. Of course, since we have shown that $\{y_t\}$ and $\{\varepsilon_t\}$ are completely equivalent, we will base our estimate on $\{\varepsilon_t\}$.

The problem we address then is, given the innovations record $\{\varepsilon_s\}_{s=0}^{t}$, find the LLS estimate $\hat{x}_{\tau/t}$ of the random variable x_t. Using a similar argument as in section 2.b, we obtain

$$\hat{x}_{\tau/t} = \sum_{s=0}^{t} R_{x\varepsilon}(\tau,s)\, R_{\varepsilon}^{-1}(\tau,\tau)\, \varepsilon_{\tau} \qquad (2.14)$$

Recursive equations for $R_{x\varepsilon}(t,\tau)$ and $R_{\varepsilon}(\tau,\tau)$ are easily obtained as

$$R_{x\varepsilon}(t,\tau) = E\{x_t\,(y_{\tau} - \hat{y}_{\tau/\tau-1})^T\}$$

$$= R_{xy}(t,\tau) - E\{x_t(\sum_{s=0}^{\tau-1} R_{y\varepsilon}(\tau,s)\, R_{\varepsilon}^{-1}(s,s)\, \varepsilon_s)^T\}$$

$$R_{x\varepsilon}(t,\tau) = R_{xy}(t,\tau) - \sum_{s=0}^{\tau-1} R_{x\varepsilon}(t,s)\, R^{-1}(s,s)\, R_{y\varepsilon}^{T}(\tau,s) \qquad (2.15)$$

Equations (2.14) and (2.15) give a completely recursive solution to the problem of obtaining a LLS $\hat{x}_{\tau/t}$. Again, as in section 2.b, we can comment that this solution is in general not finitely recursive.

The estimation filter represented by equation (2.14) will be referred to as the general innovations filter, GIF. This representation of the estimate $\hat{x}_{\tau/t}$ will form the basis of all our developments in the next two sections. The structure of the estimator is depicted in figure 2.3. Let us comment here that the estimate $\hat{x}_{\tau/t}$ is called the filtered estimate when $\tau = t$, the predicted estimate when $\tau > t$, and the smoothed estimate for $\tau < t$.

As an application let us point out that for $\{x_t = y_t\}$, we obtain the k-step ahead predicted estimate $\hat{y}_{t/t-k}$ of y_t $(t > k)$ from (2.14) as

$$\hat{y}_{t/t-k} = \sum_{\tau=0}^{t-k} R_{y\varepsilon}(t,\tau)\, R_{\varepsilon}^{-1}(\tau,\tau)\, \varepsilon_\tau \qquad (2.16)$$

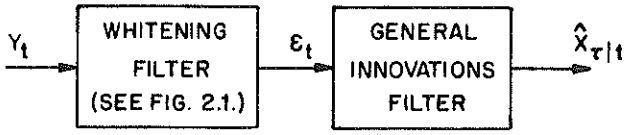

Fig. 2.3. Obtaining the estimates $\hat{x}_{\tau/t}$ from the observations $\{y_s\}$ by passing the "whitened" observations through the GIF.

Finally, it should be noted that, although $\hat{x}_{\tau/t}$ as given by equation (2.14) is the best linear estimate in mean square sense, it is in general not the best absolutely. The least squares estimate is given by the conditional expectation $\hat{x}^*_{\tau/t} = E\{ x_\tau | y_s \;,\; 0 \leq s \leq t \}$ (see e.g. [16]). In general $\hat{x}^*_{\tau/t}$ will be a nonlinear function of the data, and the complete joint probability density function of the processes $\{x_t\}$ and $\{y_t\}$ must be known in order to represent it. In the case that, as we assume here, only

covariance information is given, the LLS is the best we can do. In any case it can be shown (see also [16]) that when $\{y_t\}$ and $\{x_t\}$ are jointly Gaussian the LLS estimate is also "overall" the best. In this case, the innovations process $\{\varepsilon_t\}$ will be a white Gaussian noise (WGN).