CHAPTER 7

Recursive Estimation and the Kalman Filter

The concept of least-squares regression originates with two people. It is nowadays accepted that Legendre (1752–1833) was responsible for the first published account of the theory in 1805; and it was he who coined the term *Moindes Carrés* or least squares [6]. However, it was Gauss (1777–1855) who developed the method as a statistical tool by embedding it in a context which involved a probabilistic treatment of errors of observation. Confusion over the rival claims of priority arises from the fact that, although his first published exposition of the method appeared in 1809 in *Theoria Motus Corporum Celestium* [2], when he was 31 years of age, Gauss claimed that he had formulated his ideas many years earlier when he was in his early twenties. These matters are dealt with in Stigler's book on the History of Statistics [8].

The first exposition of the method of least squares by Gauss, which is to be found in *Theoria Motus*, is in connection with the estimation of the six coefficients which determine the elliptical orbit of a planetary body when the available observations exceed the number of parameters. His second exposition was presented in a series of papers from 1821, 1823 and 1826 which were collected together under the title *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae* [3]. It was in these papers that Gauss presented the famous theorem that *amongst all linear unbiased estimators, the least-squares estimator has minimum mean-square error.* This is know nowadays as the Gauss–Markov theorem.

The relevance of Gauss's second exposition to the theory of recursive least-squares estimation and to the concept of the Kalman filter lies in a brief passage where Gauss shows that it is possible *to find the changes which the most likely values of the unknowns undergo when a new equation is adjoined, and to determine the weights of these new determinations.* This passage refers to the business of augmenting the normal equations when a new observation becomes available. In effect, Gauss developed the algorithm of recursive least-squares estimation.

Gauss's algorithm for recursive least-squares estimation was ignored for almost a century and a half before it was rediscovered on two separate occasions. The first rediscovery was by Plackett [7] in 1950, which was before the advent of efficient on-line electronic computing; and this also passed almost unnoticed. It was the second rediscovery of the recursive algorithms in 1960 in the context of control theory which was the cue to a rapid growth of interest. Stemming from the papers of Kalman [4] and Kalman and Bucy [5] a vast literature on Kalman filtering has since accumulated.

Plackett's exposition of the recursive least-squares algorithm is within an algebraic framework which invokes only the statistical concepts of the classical linear regression model. Kalman's derivation was within the wider context of a state-space model with time-varying parameters. Although the core of the Kalman filter is still the Gauss–Plackett algorithm of recursive least-squares estimation, the widening of the context adds significantly to the extent and to the complexity of the algebra.

It seems certain that Kalman was unaware of the contributions of Gauss and Plackett; and his techniques of deriving the algorithm were quite different from theirs. He based his derivation upon the use of orthogonal projectors in deriving the minimum-mean-square-error predictors. His derivation invokes the concept of an infinite-dimensional Hilbert space.

Since Kalman's seminal paper, several other derivations have been offered, and a welter of alternative notation has arisen. Most of the alternative derivations attempt to avoid the concepts of Hilbert space and to reduce the terminology of the derivation to something closer to that of the ordinary theory of least-squares regression. Other derivations have been from a maximum-likelihood or a Bayesian standpoint. The derivation which has attracted the attention of econometricians is that of Duncan and Horn [1]. This exploits the concept of mixed estimation which originates with Theil and Goldberger [9] and which was extended by Theil [10]. An account of the method is to be found in the textbook of Theil [11, 347–352].

The method of mixed estimation is often derided by Bayesian theorists who describe it as back-door Bayesianism. In their view, it represents an attempt to use Bayesian methods without espousing the relevant Bayesian concepts, To be fair, it must be said that Theil has provided, in his textbook, an account of the Bayesian interpretation of the mixed estimation technique [11, 670–672]; and this provides an excellent way of understanding the paradigm shift which is involved in passing from classical concepts to Bayesian concepts.

The derivation of the Kalman filter by Duncan and Horn [1], although based in familiar territory, is, to my mind, utterly confusing. Its only virtue is that the notation seems familiar. The state vector, whose estimation is the object of the exercise, is compared with the vector of regression parameters in a classical linear model. However, it is nowhere clear whether this vector is to be regarded as constant or as random. This makes the crucial concept of the dispersion matrix associated with the vector particularly confusing.

It seems that, in order to clarify the statistical issues which are entailed by the Kalman filter, one must adhere rigorously either to classical concepts or to Bayesian concepts. To mix the two is a recipe for confusion unless one has a facility for passing from one to the other with ease. The derivation of the Kalman filter is not a good context in which to acquire such a facility.

The essential truth about the Kalman filter is that it is enormously complex. Its derivation, by whatever method, is bound to be lengthy and its equations are difficult to memorise. It is precisely this complexity which gives the Kalman filter its enormous power. It represents an *omnium gatherum* for a wide range of problems in statistical inference.

A comparison with the difficulties of the theory of quantum mechanics might be in order. Quantum theory has been effective is solving a wide range of problems in physics, chemistry and electronics. Nevertheless, the philosophical foundations of the theory have been a matter of debate ever since its emergence in the 1930's. Most practitioners of the quantum mechanics would agreed that, if one keeps ones eyes to the ground and watches each step, then one can proceed without undue difficulty. It is when one begins to contemplate the larger issues of meaning and of methodology that one runs the danger of falling into a vicious confusion.

Conditional Expectations and Classical Regression Theory

In deriving the algorithm of recursive least-squares estimation and in generalising it to obtain the Kalman filter, we shall rely upon the calculus of conditional expectations. Our approach is one which might be described as covert Bayesianism as distinct from back-door Bayesianism. That is to say, the derivation will be compatible with the principles of Bayesian inference, albeit that few of the Bayesian concepts will be invoked.

The calculus of conditional expectations can be derived within the context of a simple regression model which is classical in the sense that a regression relationship is postulated in which the unknown parameters are regarded as fixed quantities. We shall uncover some essential relationships within the underlying theoretical regression relationship; and, in order to obtain empirical estimators, we shall invoke the method of moments.

The method of moments is the principle of estimation which declares that, in order to derive consistent estimators of population parameters, we need only replace the theoretical moments within the set of relationships which determine these parameters by the corresponding sample moments.

Let x and y be random vectors whose joint distribution is characterised by well-defined first and second-order moments. In particular, let us define the following second-order moments of x and y

$$\begin{aligned} D(x) &= E(xx') - E(x)E(x'), \\ D(y) &= E(yy') - E(y)E(y'), \\ C(y,x) &= E(yx') - E(y)E(x'). \end{aligned} \tag{1}$$

Also, let us postulate that the conditional expectation of y given x is a simple linear function of x:

$$E(y|x) = \alpha + B'x. \tag{2}$$

Then the object is to find expressions for the vector α and the matrix B which are in terms of the moments listed under (1).

We begin by multiplying $E(y|x)$ by the marginal density function of x and by integrating with respect to x. This converts the conditional expectation into an unconditional expectation. The general result may be expressed by writing

$$E\big\{E(y|x)\big\} = E(y). \tag{3}$$

On applying the latter to equation (2), we find that

$$E(y) = \alpha + B'E(x), \qquad \text{or} \qquad \alpha = E(y) - B'E(x). \tag{4}$$

Next, by multiplying $E(y|x)$ by x' and by the marginal marginal density function of x, and by integrating with respect to x, we obtain the joint moment $E(xy')$. Thus, from equation (2), we get

$$E(yx') = \alpha E(x) + B'E(xx'). \tag{5}$$

But, postmultiplying the first equation under (4) by $E(x')$ gives

$$E(y)E(x') = \alpha E(x') + B'E(x)E(x'), \tag{6}$$

and, when this is subtracted from (5), the result, in view of the definitions under (1), is

$$\begin{aligned} C(y,x) &= E(yx') - E(y)E(x') \\ &= B'\big\{E(xx') - E(x)E(x')\big\} \\ &= B'D(x). \end{aligned} \tag{7}$$

The result from (7) is that

$$B' = C(y,x)D^{-1}(x). \tag{8}$$

This expression for B and the expression for α under (4) can be substituted into equation (2) to give

$$\begin{aligned} E(y|x) &= \alpha + Bx \\ &= E(y) - B'E(x) + B'x \\ &= E(y) + C(y,x)D^{-1}(x)\big\{x - E(x)\big\}. \end{aligned} \tag{9}$$

In the usual presentation of the theory of the classical regression model, the observations on x and y for $t = 1,\ldots,T$ are accumulated in the matrices X and Y as successions of row vectors, each arrayed below its predecessor. If the matrices X and Y contain the mean-adjusted observations, then the products $T^{-1}X'X$ and $T^{-1}X'Y$ become the empirical counterparts of the moment matrices $D(x)$ and $C(x,y)$ respectively. The estimator of B derived from the principle of the method of moments is $\hat{B} = (X'X)^{-1}X'Y$.

Several additional results in the algebra of conditional expectations which we shall invoke in the next section can also be derived with ease. To avoid burdening this account with unnecessary developments, let us simply declare in summary that, if x, y are jointly distributed variables which bear the linear relationship $E(y|x) = \alpha + B'x$, then

$$E(y|x) = E(y) + C(y,x)D^{-1}(x)\big\{x - E(x)\big\}, \tag{10}$$

$$D(y|x) = D(y) - C(y,x)D^{-1}(x)C(x,y), \tag{11}$$

$$E\big\{E(y|x)\big\} = E(y), \tag{12}$$

$$D\big\{E(y|x)\big\} = C(y,x)D^{-1}(x)C(x,y), \tag{13}$$

$$D(y) = D(y|x) + D\big\{E(y|x)\big\}, \tag{14}$$

$$C\big\{y - E(y|x), x\big\} = 0. \tag{15}$$

Recursive Least-Squares Estimation

We may use the results in the algebra of conditional expectations presented above to derive the algorithm for the recursive least-squares estimation of the parameters of a classical linear regression model. The tth instance of the regression relationship is represented by

$$y_t = x_t'\beta + \varepsilon_t. \tag{16}$$

Here y_t is a scalar element instead of the vector which appears in equation (2). Notice also that the expression for the mean value has undergone a transposition so that we have $x_t'\beta$ instead of $\beta' x_t$. Since the mean value is also a scalar, nothing is affected. It is assumed that the disturbances ε_t are serially independent with

$$E(\varepsilon_t) = 0 \quad \text{and} \quad V(\varepsilon_t) = \sigma^2 \quad \text{for all} \quad t. \tag{17}$$

In order to initiate the recursion, there must be an initial estimate b_0 of β together with a corresponding dispersion matrix. In the usual context of classical regression theory, we should regard this dispersion matrix as the variance–covariance matrix of the estimator. Instead, we are inclined to attribute a distribution to β and to regard $b_0 = E(\beta)$ and $P_0 = D(\beta)$ as its mean and its dispersion matrix. This distribution is, in effect, a Bayesian prior.

The empirical information available at time t is the set of observations $\mathcal{I}_t = \{y_1, \ldots, y_t\}$. In an alternative notation, we would use $\mathcal{I}_0 = \{\beta_0, P_0\}$ to denote the prior information which would be included together with the empirical information in all information sets $\mathcal{I}_t$ with $t > 0$.

Our object is to derive the estimates $b_t = E(\beta|\mathcal{I}_t)$ and $P_t = D(\beta|\mathcal{I}_t)$ from the information available at time t in a manner which makes best use of the previous estimates $b_{t-1} = E(\beta|\mathcal{I}_{t-1})$ and $P_{t-1} = D(\beta|\mathcal{I}_{t-1})$. The first task is to evaluate the expression

$$E(\beta|\mathcal{I}_t) = E(\beta|\mathcal{I}_{t-1}) + C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\big\{y_t - E(y_t|\mathcal{I}_{t-1})\big\}, \tag{18}$$

which is derived directly from (10). There are three elements on the RHS which require further development. The first is the term

$$\begin{aligned} y_t - E(y_t|\mathcal{I}_{t-1}) &= y_t - x_t' b_{t-1} \\ &= h_t. \end{aligned} \tag{19}$$

This is the error from predicting y_t from the information available at time $t-1$. Next is the dispersion matrix of associated with this prediction. This is

$$\begin{aligned} D(y_t|\mathcal{I}_{t-1}) &= D\big\{x_t'(\beta - b_{t|t-1})\big\} + D(\varepsilon_t) \\ &= x_t' P_{t-1} x_t + \sigma^2 = D(h_t). \end{aligned} \tag{20}$$

Finally there is the covariance

$$\begin{aligned} C(\beta, y_t|\mathcal{I}_{t-1}) &= E\big\{(\beta - b_{t-1})y_t'\big\} \\ &= E\big\{(\beta - b_{t-1})(x_t'\beta + \varepsilon_t)'\big\} \\ &= P_{t-1}x_t. \end{aligned} \tag{21}$$

On putting these elements together, we get

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + \sigma^2)^{-1}(y_t - x_t'b_{t-1}). \tag{22}$$

There must also be a means of deriving the dispersion matrix $D(\beta|\mathcal{I}_t) = P_t$ from its predecessor $D(\beta|\mathcal{I}_{t-1}) = P_{t-1}$. Equation (11) indicates that

$$D(\beta|\mathcal{I}_t) = D(\beta|\mathcal{I}_{t-1}) - C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})C(y_t, \beta|\mathcal{I}_{t-1}). \tag{23}$$

It follows from (20) and (21) that this is

$$P_t = P_{t-1} - P_{t-1}x_t(x_t'P_{t-1}x_t + \sigma^2)^{-1}x_t'P_{t-1}. \tag{24}$$

It is useful, for future reference, to anatomise the components of the recursive least-squares algorithm. A summary of the equations, which entails some further definitions, is as follows:

$$h_t = y_t - x_t'b_{t-1}, \qquad \textit{Prediction Error} \tag{25}$$

$$f_t = x_t'P_{t-1}x_t + \sigma^2, \qquad \textit{Error Dispersion} \tag{26}$$

$$\kappa_t = P_{t-1}x_t f_t^{-1}, \qquad \textit{Filter Gain} \tag{27}$$

$$b_t = b_{t-1} + \kappa_t h_t, \qquad \textit{Parameter Estimate} \tag{28}$$

$$P_t = (I - \kappa_t x_t)P_{t-1}. \qquad \textit{Estimate Dispersion} \tag{29}$$

Alternative expressions are available for P_t and κ_t:

$$P_t = (P_{t-1}^{-1} + \sigma^{-2}x_t x_t')^{-1}, \tag{30}$$

$$\kappa_t = \sigma^{-2}P_t x_t. \tag{31}$$

The expression on the RHS of (30) is confirmed by using the well-known matrix inversion formula

$$(B + CDC')^{-1} = B^{-1} - B^{-1}C(C'B^{-1}C + D^{-1})^{-1}C'B^{-1} \tag{32}$$

to recover the original expression for P_t given under (29). To verify the identity $P_{t-1}x_t f_t^{-1} = P_t x_t \sigma^{-2}$ which equates (27) and (31), we write it as $P_t^{-1}P_{t-1}x_t = x_t\sigma^{-2}f_t$. The latter is readily confirmed using the expression for P_t from (30) and the expression for f_t from (26).

Equation (30) indicates that

$$\sigma^2 P_t^{-1} = \sigma^2 P_0^{-1} + \sum_{i=1}^{t} x_i x_i'. \tag{33}$$

Apart from the matrix $\sigma^2 P_0^{-1}$, which becomes relatively insignificant for large values of t, this is just the familiar moment matrix of ordinary least-squares regression.

When equations (30) and (31) are used in (28), we get the following expression for recursive least-squares estimate:

$$b_t = b_{t-1} + \sigma^{-2}(P_{t-1}^{-1} + \sigma^{-2}x_t x_t')^{-1}x_t(y_t - x_t'b_{t-1}). \tag{34}$$

The equation serves to show that σ^2, which is a factor of P_t, can be cancelled from the formula for b_t.

The formula of (34) certainly appears to be simpler than that of (22). However, in comparison to the latter, it is computationally inefficient. The formula of (22) entails finding the inverse of the scalar element $f_t = x_t P_{t-1} x_t' + \sigma^2$ which represents the dispersion of the prediction error. The formula under (33) involves the inversion of the entire matrix P_t. To use this formula in place of that of (22) would be to loose all the computational advantages of the recursive least-squares algorithm.

Extensions of the Recursive Least-Squares Algorithm

The algorithm which we have presented in the previous section represents little more than an alternative means of computing the ordinary least-squares regression estimates. If the parameters of the underlying process which generates the data are stable, then we can expect the estimate b_t to converge also to a stable value as the number of observations t increases At the same time, the elements of the dispersion matrix P_t will decrease in value.

A further consequence of the growth of the number of observations is that the filter gain κ_t will diminish at t increases. This implies that the impact of successive prediction errors upon the estimate of β will diminish as the amount of information already incorporated in the estimate increases.

If there is doubt about the constancy of the regression parameter, then it may be desirable to give greater weight to the more recent data; and it might even be appropriate to discard data which has reached a certain age and has passed its date of expiry.

One way of accommodating parametric variability is to base the estimate on only the most recent portion of the data. As each new observation is acquired another observation may be removed so that, at any instant, the estimator comprises only n points. Such an estimator has been described as a rolling regression. Implementations are available in the recent versions of the more popular econometric computer packages such as *Microfit 3.0* and *PCGive*.

It is a simple matter to extend the algorithm of the previous section to produce a rolling regression. The additional task is to remove the data which was acquired at time $t-n$. The first step is to adjust the moment matrix to give $\sigma^2 P_t^{*-1} = \sigma^2 P_{t-1}^{-1} - x_{t-n}x_{t-n}'$. The matrix inversion formula of (32) indicates that

$$\begin{aligned} P_t^* &= (P_{t-1}^{-1} - \sigma^{-2}x_{t-n}x_{t-n}')^{-1} \\ &= P_{t-1} - P_{t-1}x_{t-n}(x_{t-n}'P_{t-1}x_{t-n} - \sigma^2)^{-1}x_{t-n}'P_{t-1}, \end{aligned} \tag{35}$$

Next, an intermediate estimate b_t^*, which is based upon the reduced information, is obtained from b_{t-1} via the formula

$$\begin{aligned} b_t^* &= b_{t-1} - \sigma^{-2}P_t^{*-1}x_{t-n}(y_{t-n} - x_{t-n}'b_{t-1}) \\ &= b_{t-1} - P_{t-1}x_{t-n}(x_{t-n}'P_{t-1}x_{t-n} - \sigma^2)^{-1}(y_{t-n} - x_{t-n}'b_{t-1}). \end{aligned} \tag{36}$$

This formula can be understood by considering the inverse problem of obtaining b_{t-1} from b_t^* by the *addition* of the information from time $t-n$. A rearrangement of the resulting expression for b_{t-1} gives the initial expression for b_t^* under (36). Finally, the estimate b_t, which is based on the n data points $x_t, \ldots, x_{t-n+1}$, is obtained from the formula under (22) by replacing b_{t-1} with b_t^* and P_{t-1} with P_t^*.

Discarding observations which have passed a date of expiry is an appropriate procedure when the processes generating the data are liable, from time to time, to undergo sudden structural changes. For it ensures that any misinformation which is conveyed by the data which predate the structural change will not be kept on record permanently. However, if the processes are expected to change gradually in a more or less systematic fashion, then a gradual discounting of old data may be more appropriate. An exponential weighting scheme applied to the data might serve this purpose.

Let the rate at which the data is discounted be given by a parameter $\lambda \in (0, 1]$. Then, in place of the expression for P_t under (30), we should have

$$\begin{aligned} P_t &= (\lambda P_{t-1}^{-1} + \sigma^{-2} x_t x_t')^{-1} \\ &= \frac{1}{\lambda}\Big\{ P_{t-1} - P_{t-1} x_t (x_t' P_{t-1} x_t + \lambda\sigma^2)^{-1} x_t P_{t-1} \Big\}. \end{aligned} \tag{37}$$

The formula for the parameter estimate would be

$$b_t = b_{t-1} + P_{t-1} x_t (x_t' P_{t-1} x_t + \lambda\sigma^2)^{-1} (y - x_t' b_{t-1}). \tag{38}$$

It is curious that econometric packages mentioned above have implemented rolling regression but not exponentially-weighted regression.

A wide variety of techniques for shaping the memory of the recursive least-square algorithm may be devised. However, it is clear that such formulations are essentially pragmatic, and one might wish for a theoretical basis from which to develop the algorithms. The basis is provided by the fully-fledged Kalman filter.

The elaboration of the recursive least-square model which is required in order to achieve the generality of the Kalman filter is the addition of a process which describes the variation of the parameter vector β. Such a process might be described by the equation

$$\beta_t = \Phi \beta_{t-1} + \nu_t, \tag{39}$$

which represents a Markov scheme. We shall consider such an elaboration in the next section. However, we shall begin by adopting a new notation. The reason is that the Kalman filter is a system which accommodates a very wide range of models; and one should avoid making references automatically to the regression model.

Equations of the Kalman Filter

We shall present the basic equations of the Kalman filter in the briefest possible manner. The state-space model, which underlies the Kalman filter, consists of two equations

$$y_t = H_t \xi_t + \eta_t, \qquad \textit{Observation Equation} \tag{40}$$

$$\xi_t = \Phi_t \xi_{t-1} + \nu_t, \qquad \textit{Transition Equation} \tag{41}$$

where y_t is the observation on the system and ξ_t is the state vector. The observation error η_t and the state disturbance ν_t are mutually uncorrelated random vectors of zero mean with dispersion matrices

$$D(\eta_t) = \Omega_t \qquad \text{and} \qquad D(\nu_t) = \Psi_t. \tag{42}$$

The observation equation is analogous to the regression equation of (16), whereas the transition equation is simply (39) in new notation.

It is assumed that the matrices H_t, Φ_t, Ω_t and Ψ_t are known for all $t = 1, \ldots, n$ and that an initial estimate x_0 is available for the state vector ξ_0 at time $t = 0$ together with a dispersion matrix $D(\xi_0) = P_0$. The empirical information available at time t is the set of observations $\mathcal{I}_t = \{y_1, \ldots, y_t\}$.

The Kalman-filter equations determine the state-vector estimates $x_{t|t-1} = E(\xi_t|\mathcal{I}_{t-1})$ and $x_t = E(\xi_t|\mathcal{I}_t)$ and their associated dispersion matrices $P_{t|t-1}$ and P_t. From $x_{t|t-1}$, the prediction $\hat{y}_{t|t-1} = H_t x_{t|t-1}$ is formed which has a dispersion matrix F_t. A summary of these equations is as follows:

$$x_{t|t-1} = \Phi_t x_{t-1}, \qquad \textit{State Prediction} \tag{43}$$

$$P_{t|t-1} = \Phi_t P_{t-1} \Phi_t' + \Psi_t, \qquad \textit{Prediction Dispersion} \tag{44}$$

$$e_t = y_t - H_t x_{t|t-1}, \qquad \textit{Prediction Error} \tag{45}$$

$$F_t = H_t P_{t|t-1} H_t' + \Omega_t, \qquad \textit{Error Dispersion} \tag{46}$$

$$K_t = P_{t|t-1} H_t' F_t^{-1}, \qquad \textit{Kalman Gain} \tag{47}$$

$$x_t = x_{t|t-1} + K_t e_t, \qquad \textit{State Estimate} \tag{48}$$

$$P_t = (I - K_t H_t) P_{t|t-1}. \qquad \textit{Estimate Dispersion} \tag{49}$$

In comparison with the equations of the recursive regression algorithm listed under (25)–(22), there are two additions: equation (43) for the state prediction and equation (44) for its dispersion. These owe their existence to the presence of the transition equation (41); and they vanish when $\Phi = I$ and $\nu_t = 0$.

Alternative expressions are available for P_t and K_t on the assumption that Ω_t is nonsingular:

$$P_t = (P_{t|t-1}^{-1} + H_t' \Omega_t^{-1} H_t)^{-1}, \tag{50}$$

$$K_t = P_t H_t' \Omega_t^{-1}. \tag{51}$$

By applying the matrix inversion lemma to the expression on the RHS of (50), we obtain the original expression for P_t given under (49). To verify the identity $P_{t|t-1} H_t' F_t^{-1} = P_t H_t' \Omega_t^{-1}$ which equates (47) and (51), we write it as $P_t^{-1} P_{t|t-1} H_t' = H_t' \Omega_t^{-1} F_t$. The latter is readily confirmed using the expression for P_t from (50) and the expression for F_t from (46).

The equations of the Kalman filter may be derived using the results from the algebra of conditional expectations which are listed under (10)–(15).

Of the equations listed under (43)—(49), those under (45) and (47) are merely definitions.

To demonstrate equation (43), we use (12) to show that

$$
\begin{aligned}
E(\xi_t|\mathcal{I}_{t-1}) &= E\big\{E(\xi_t|\xi_{t-1})|\mathcal{I}_{t-1}\big\}\\
&= E\big\{\Phi_t\xi_{t-1}|\mathcal{I}_{t-1}\big\}\\
&= \Phi_t x_{t-1}.
\end{aligned}
\tag{52}
$$

We use (14) to demonstrate equation (44):

$$
\begin{aligned}
D(\xi_t|\mathcal{I}_{t-1}) &= D(\xi_t|\xi_{t-1}) + D\big\{E(\xi_t|\xi_{t-1})|\mathcal{I}_{t-1}\big\}\\
&= \Psi_t + D\big\{\Phi_t\xi_{t-1}|\mathcal{I}_{t-1}\big\}\\
&= \Psi_t + \Phi_t P_{t-1}\Phi_t'.
\end{aligned}
\tag{53}
$$

To obtain equation (46), we substitute (40) into (45) to give $e_t = H_t(\xi_t - x_{t|t-1}) + \eta_t$. Then, in view of the statistical independence of the terms on the RHS, we have

$$
\begin{aligned}
D(e_t) &= D\big\{H_t(\xi_t - x_{t|t-1})\big\} + D(\eta_t)\\
&= H_t P_{t|t-1} H_t' + \Omega_t = D(y_t|\mathcal{I}_{t-1}).
\end{aligned}
\tag{54}
$$

To demonstrate the updating equation (48), we begin by noting that

$$
\begin{aligned}
C(\xi_t, y_t|\mathcal{I}_{t-1}) &= E\big\{(\xi_t - x_{t|t-1})y_t'\big\}\\
&= E\big\{(\xi_t - x_{t|t-1})(H_t\xi_t + \eta_t)'\big\}\\
&= P_{t|t-1}H_t'.
\end{aligned}
\tag{55}
$$

It follows from (10) that

$$
\begin{aligned}
E(\xi_t|\mathcal{I}_t) &= E(\xi_t|\mathcal{I}_{t-1}) + C(\xi_t, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\big\{y_t - E(y_t|\mathcal{I}_{t-1})\big\}\\
&= x_{t|t-1} + P_{t|t-1}H_t'F_t^{-1}e_t.
\end{aligned}
\tag{56}
$$

The dispersion matrix under (49) for the updated estimate is obtained via equation (11):

$$
\begin{aligned}
D(\xi_t|\mathcal{I}_t) &= D(\xi_t|\mathcal{I}_{t-1}) - C(\xi_t, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})C(y_t, \xi_t|\mathcal{I}_{t-1})\\
&= P_{t|t-1} - P_{t|t-1}H_t'F_t^{-1}H_tP_{t|t-1}.
\end{aligned}
\tag{57}
$$

References

[1] Duncan, D.B. and S.D. Horn, (1972), Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis, Journal of the American Statistical Association, 67, 815–821

[2] Gauss, K.F., (1809), Theoria Motus Corporum Celestium.

[3] Gauss, K.F., (1823), Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.

[4] Kalman, R.E., (1960), A New Approach to Linear Filtering and Prediction Problems, Trans. ASME, J. Basic Eng., 82, 35–45.

[5] Kalman, R.E. and R.S. Bucy, (1961), New Results in Linear Filtering and Prediction Theory, Trans. ASME, J. Basic Eng., 83, 95–107

[6] Legendre, A.M., (1805), Nouvelles Méthodes pour la Determination des Orbites des Comètes.

[7] Plackett, R.L., (1950), Some Theorems in Least Squares, Biometrika, 37, 149–157.

[8] Stigler, S.M., (1986), The History of Statistics, Harvard University Press: Cambridge, Mass.

[9] Theil, H. and A.S. Goldberger, (1961), On Pure and Mixed Statistical Estimation in Economics, International Economic Review, 2, 65–78

[10] Theil, H., (1971), Principles of Econometrics, John Wiley and Sons, New York.

[11] Theil, H., (1963), On the Use of Incomplete Prior Information in Regression Analysis, Journal of the American Statistical Association, 58, 401–414.