Recursive Estimation in Econometrics

D.S.G. Pollock

Department of Economics, Queen Mary College, University of London, Mile End Road, London E1 4NS, UK

Abstract

An account is given of recursive regression and Kalman filtering that gathers the important results and the ideas that lie behind them. It emphasises areas where econometricians have made contributions, including methods for handling the initial-value problem associated with nonstationary processes and algorithms for fixed-interval smoothing.

Key words: Recursive Regression, Kalman Filtering, Fixed-Interval Smoothing, The Initial-Value Problem

1 Introduction

The algorithms for recursive estimation and Kalman filtering are being used increasingly in applied econometrics, but econometricians have been slower than other statisticians to exploit them. The second section of the paper describes how the use has developed.

The third section lays essential groundwork by expounding the algorithm for ordinary recursive regression. This provides a preparation for the complexities of the Kalman filter, whose features are more easily understood when related to something similar but simpler.

The treatment given recursive regression in Sections 3 and 4 has a Bayesian flavour and relies on the calculus of conditional expectations, whose essentials are provided in an appendix.

Section 5 examines the prediction-error decompositions associated with recursive regression, whilst Section 6 deals with extensions and elaborations of recursive regression and describes some applications in control engineering that can be exploited by econometricians.

Section 7, treats the Kalman filter, depicted as an elaboration of the preceding regression algorithm. The next two sections deal with the likelihood

Email address: d.s.g.pollock@qmw.ac.uk (D.S.G. Pollock).

function and the starting-value problem. The smoothing operations described in Section 10 take account of this problem.

An extensive bibliography contains references to the work of econometricians on recursive estimation and the sources on which they have relied. Because of the complexity and diversity of the notation, readers of this material are advised to maintain a glossary to assist in making the necessary translations and comparisons.

Many contributions to the literature on Kalman filtering assume familiarity with the algebra. Those by econometricians have come in small increments through long sequences of papers that often refer only to their immediate predecessors. Seldom do they recapitulate the original motivations. Such literature makes for difficult reading. One of the purposes of this paper is to make the important results and the ideas that lie behind them more accessible by gathering them in one place.

2 Historical Aspects

Least-squares regression originates with two people. Legendre (1805) gave the first published account of the theory and coined the term *Moindres Carrés* or least squares. However, Gauss developed the method as a statistical tool by giving the errors a probabilistic treatment. Confusion over priority arises because Gauss claimed that he had formulated his ideas many years before his first published exposition of the method, which appeared in 1809 in *Theoria Motus Corporum Celestium*. These matters are dealt with in the book of Stigler (1986) on the History of Statistics.

Gauss's first exposition of the method of least squares in *Theoria Motus* deals with the estimation of the six coefficients that 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, collected together under the title *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. It was here that Gauss presented the famous theorem, now known as the Gauss–Markov theorem, that *amongst all linear unbiased estimators, the least-squares estimator has minimum mean-square error.*

The relevance of Gauss's second exposition to recursive least-squares estimation and the Kalman filter lies in a brief passage where he 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 refers to a method of augmenting the normal equations with new observations which is, effectively, the algorithm of recursive least-squares estimation. The French translation of this passage, due to Bertrand (1855), is reproduced by Young (1984) in an appendix accompanied by a synoptic commentary that interprets the results in a modern notation.

Gauss's algorithm for recursive least-squares estimation was ignored for

almost a century and a half before it was rediscovered twice. The first rediscovery was in Plackett (1950), before the advent of efficient on-line electronic computing. This passed almost unnoticed. The second rediscovery was in 1960 in the context of control theory; and this was a spur to a rapid growth of interest. Stemming from Kalman (1960) and Kalman and Bucy (1961), a vast literature on Kalman filtering has accumulated.

Plackett's exposition of recursive least-squares invokes only the statistical concepts of the classical linear regression model. Kalman's derivation is within the context of a state-space model with time-varying parameters. Although the core of the Kalman filter is still the Gauss–Plackett algorithm, widening the context greatly increases the extent and complexity of the algebra.

It seems certain that Kalman was unaware of the contributions of Gauss and Plackett. His techniques for deriving the algorithm are quite different from theirs. He uses orthogonal projectors in deriving the minimum-mean-square-error predictors within an infinite-dimensional Hilbert space.

Since Kalman's seminal paper, several other derivations have been offered, creating a welter of alternative notation. Most avoid Hilbert spaces and use terminology closer to that of ordinary least-squares regression. Others adopt a maximum-likelihood or a Bayesian standpoint.

The derivation that first attracted the attention of econometricians is in Duncan and Horn (1972). It exploits the concept of mixed estimation developed in Theil and Goldberger (1961) and extended in Theil (1963). An account of this method is found in Theil (1971, pps. 347–352). Recent accounts adopt a Bayesian approach, as in Durbin and Koopman (2001).

The slowness of econometricians in adopting the Kalman filter may reflect their reluctance to espouse time-varying parameters. They have tended to assume that, instead of flexing or bending, their structural models will break at identifiable points. As we shall describe in Sections 5 and 6, recursive regression is being used increasingly in detecting such breaks.

The principal econometric uses of the Kalman filter and the associated fixed-interval smoothing algorithms, have been in trend estimation and signal extraction, of which there is now a considerable literature. Harrison and Stevens (1976), which foreshadows the development of structural time series models, has been highly influential here as have Harvey and Todd (1983) and Gersch and Kitigawa (1983) and the book of Harvey (1989).

An equally influential alternative methodology, implemented by means other than the Kalman filter, such as the method of Burman (1980), is found in Cleveland and Tiao (1976), Hillmer and Tiao (1982) and Maravall (1985). Much of the relevant literature is cited in Pollock (2000, 2001a, 2001b, 2002), where alternatives to the Kalman filter are employed.

Another growing use of the Kalman filter is as a device for calculating the likelihood functions of time series models when estimating their parameters. After a model is represented in state-space form, the likelihood function can be evaluated via the prediction-error decomposition, as was demonstrated originally in Schweppe (1965).

Early econometric examples include the algorithms for evaluating the likelihood of autoregressive moving-average (ARMA) models as given in Gardner, Harvey and Phillips (1980) and Mélard (1983). Jones (1980) uses this approach for fitting ARMA models to time series with missing observations. Several state-space representations for ARMA models are described in Pollock (1999). However, current applications of this method of evaluating the likelihood function extend far beyond classical univariate time series models.

The growing econometric use of the Kalman filter and other recursive algorithms has encouraged the development of relevant software such as *SsfPack* described in Koopman, Shephard and Doornik (1999), and that provided in Bomhoff (1994).

The scientific community is now well served by freely available resources relating to the Kalman filter. An excellent starting point is the Website of Welch and Bishop ⟨`http://www.cs.unc.edu/~welch/kalman`⟩.

3 Recursive Regression

We may use the theory of conditional expectations in the appendix to derive the algorithm for recursive estimation of the classical linear regression model. The tth instance of the regression relationship is

$$y_t = x_t'\beta + \varepsilon_t, \tag{1}$$

where y_t is a scalar value and x_t is a vector of k elements. The disturbances ε_t are assumed to be serially independent and normally distributed with

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

To initiate the recursion, one needs an initial estimate of β and its dispersion matrix. In classical regression theory, this dispersion matrix is regarded as the variance–covariance matrix of the estimator. Here, we attribute a distribution to β with a mean $b_0 = E(\beta)$ and a dispersion matrix $\sigma^2 P_0 = D(\beta)$. This is, in effect, a Bayesian prior.

The information $\mathcal{I}_t$ available at time t comprises the observations and $\mathcal{I}_0$, which is $\{b_0, \sigma^2 P_0\}$, if there is prior information, and the emptyset in the absence of such information. Thus, $\mathcal{I}_t = \{y_t, \mathcal{I}_{t-1}\} = \{y_t, \ldots, y_1, \mathcal{I}_0\}$. Initially, we assume that the prior for β is fully specified, giving rise to a marginal distribution $N(y_1; \mathcal{I}_0)$ and to a sequence of conditional distributions $N(y_t|\mathcal{I}_{t-1}); t = 2, \ldots, T$, each of which presupposes its predecessors.

Our object is to derive the estimates $b_t = E(\beta|\mathcal{I}_t)$ and $\sigma^2 P_t = D(\beta|\mathcal{I}_t)$ from the information available at time t making the best use of the previous estimates $b_{t-1} = E(\beta|\mathcal{I}_{t-1})$ and $\sigma^2 P_{t-1} = D(\beta|\mathcal{I}_{t-1})$. First we must to evaluate

$$E(\beta|\mathcal{I}_t) = E(\beta|\mathcal{I}_{t-1}) + C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\{y_t - E(y_t|\mathcal{I}_{t-1})\}, \tag{3}$$

which is derived directly from (A.8.i) within the appendix. Three components on the RHS require further development. The first is

$$\begin{aligned} y_t - E(y_t|\mathcal{I}_{t-1}) &= y_t - x_t'b_{t-1} \\ &= h_t. \end{aligned} \tag{4}$$

This is the error in predicting y_t from the information available at time $t-1$.

According to (A.8.vi), the prediction error is uncorrelated with the elements of the information set $\mathcal{I}_{t-1}$. Moreover, it is independent of the previous prediction error h_{t-1}, which depends solely on the information in $\mathcal{I}_{t-1} = \{y_{t-1}, \mathcal{I}_{t-2}\}$. By reverting this argument to the start of the sample, the prediction errors are seen to form a sequence of mutually independent random variables. Moreover, given $\mathcal{I}_0 = \{b_0, \sigma^2 P_0\}$, there is a one-to-one correspondence between the observations and the prediction errors; and so the information at time t is also represented by $\mathcal{I}_t = \{h_t, \ldots, h_1, \mathcal{I}_0\}$.

The second component is the dispersion matrix associated with the prediction:

$$\begin{aligned} D(y_t|\mathcal{I}_{t-1}) &= D\{x_t'(\beta - b_{t-1})\} + D(\varepsilon_t) \\ &= \sigma^2 x_t' P_{t-1} x_t + \sigma^2 = D(h_t), \end{aligned} \tag{5}$$

and the third is the covariance

$$\begin{aligned} C(\beta, y_t|\mathcal{I}_{t-1}) &= E\{(\beta - b_{t-1})y_t'\} \\ &= E\{(\beta - b_{t-1})(x_t'\beta + \varepsilon_t)'\} \\ &= \sigma^2 P_{t-1} x_t. \end{aligned} \tag{6}$$

Employing these elements in equation (3), we get

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + 1)^{-1}(y_t - x_t'b_{t-1}). \tag{7}$$

There must also be a means for deriving the dispersion matrix $D(\beta|\mathcal{I}_t) = \sigma^2 P_t$ from its predecessor $D(\beta|\mathcal{I}_{t-1}) = \sigma^2 P_{t-1}$. Equation (A.8.ii) indicates that

$$D(\beta|\mathcal{I}_t) = D(\beta|\mathcal{I}_{t-1}) - C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})C(y_t, \beta|\mathcal{I}_{t-1}). \tag{8}$$

It follows from (5) and (6) that the desired result is

$$\sigma^2 P_t = \sigma^2 P_{t-1} - \sigma^2 P_{t-1}x_t(x_t'P_{t-1}x_t + 1)^{-1}x_t'P_{t-1}. \tag{9}$$

For future reference, we shall anatomise the components of the algorithm of recursive regression as follows:

$$h_t = y_t - x_t'b_{t-1}, \qquad \textit{Prediction Error} \tag{10}$$

$$\sigma^2 f_t = \sigma^2(x_t'P_{t-1}x_t + 1), \qquad \textit{Error Dispersion} \tag{11}$$

$$\kappa_t = P_{t-1}x_t f_t^{-1}, \qquad \textit{Filter Gain} \tag{12}$$

$$b_t = b_{t-1} + \kappa_t h_t, \qquad \textit{Parameter Estimate} \tag{13}$$

$$\sigma^2 P_t = \sigma^2 (I - \kappa_t x_t') P_{t-1}. \qquad \textit{Estimate Dispersion} \tag{14}$$

Alternative expressions are available for P_t and κ_t:

$$P_t = (P_{t-1}^{-1} + x_t x_t')^{-1}, \tag{15}$$

$$\kappa_t = P_t x_t. \tag{16}$$

To confirm (15), the matrix inversion formula of (A.3.iii) is used to recover the original expression for P_t given by (9) and (14). To verify the identity $P_{t-1}x_t f_t^{-1} = P_t x_t$, which equates (12) and (16), we write it as $P_t^{-1}P_{t-1}x_t = x_t f_t$, which is readily confirmed using the expression for P_t from (15) and the expression for f_t from (11).

Equation (15) indicates that

$$\begin{aligned} P_t^{-1} &= P_{t-1}^{-1} + x_t x_t' \\ &= P_0^{-1} + \sum_{i=1}^{t} x_i x_i'. \end{aligned} \tag{17}$$

Apart from P_0^{-1}, which becomes inconsequential when t is large, this is just the familiar moment matrix of ordinary least-squares regression.

Using (15) and (16) in (13), we get the following expression for the recursive regression estimate:

$$\begin{aligned} b_t &= b_{t-1} + (P_{t-1}^{-1} + x_t x_t')^{-1} x_t (y_t - x_t' b_{t-1}) \\ &= b_{t-1} + P_t x_t (y_t - x_t' b_{t-1}). \end{aligned} \tag{18}$$

This formula appears to be simpler than (7). However, it is computationally less efficient. Equation (7) requires the inverse of the scalar element $f_t = x_t P_{t-1} x_t' + 1$, which is the variable factor in the dispersion of the prediction error, whilst (18) requires a matrix inversion in forming P_t. Using (18) instead of (7) looses the computational advantages of the recursive regression algorithm.

However, (18) provides an opportunity for unravelling the recursive system. Multiplying the second expression for b_t by P_t^{-1} gives

$$\begin{aligned} P_t^{-1} b_t &= (P_t^{-1} - x_t x_t') b_{t-1} + x_t y_t \\ &= P_{t-1}^{-1} b_{t-1} + x_t y_t. \end{aligned} \tag{19}$$

Pursuing a recursion on the RHS and using (17) on the LHS, one finds that $(P_0^{-1} + \sum_{i=1}^{t} x_i x_i')b_t = P_0^{-1} b_0 + \sum_{i=1}^{t} x_i y_i$. Setting $t = T$ and gathering the data into $X = [x_1, \ldots, x_T]'$ and $y = [y_1, \ldots, y_T]'$ gives the equation from which the following full-sample estimator is obtained:

$$b_T = (X'X + P_0^{-1})^{-1}(X'y + P_0^{-1} b_0). \tag{20}$$

This is the so-called mixed estimator of Theil and Goldberger (1961), which is derivable by minimising the function

$$\begin{aligned} S(y,\beta) &= S(y|\beta) + S(\beta) \\ &= (y - X\beta)'(y - X\beta) + (\beta - b_0)'P^{-1}(\beta - b_0) \end{aligned} \tag{21}$$

in respect of β.

4 Initialising a Recursive Regression

In practice, when the recursive formulae are used in an ordinary regression analysis, the initial estimates of the parameter vector and their dispersion matrices are determined by an initial stretch of data. If $X_k = [x_1, \ldots, x_k]'$ is a full-rank matrix of k initial observations on the regressors and $Y_k = [y_1, \ldots, y_k]'$ contains observations on the dependent variable, then the recursion begins with $b_k = X_k^{-1}Y_k$ and $P_k = (X_k'X_k)^{-1}$. The full-sample estimator is the ordinary least-squares estimator $b = (X'X)^{-1}X'y$.

To understand the initial solution b_k, consider an arbitrarily chosen finite value b_0 with a dispersion matrix P_0 containing large diagonal elements to reflect a lack of confidence in b_0. (One might set $P_0 = \rho I$ with $\rho^{-1} \to 0$, for example.) These are so-called diffuse initial conditions. Then, if the numerical accuracy of the computer were sufficient to calculate the sequence $b_1, \ldots, b_k$ via equation (7), one would find b_k within an epsilon of $X_k^{-1}Y_k$

There are more sophisticated ways to initialise the recursive procedure, using pseudo or 'diffuse' information, that enable iterations to begin at $t = 0$. When $t = k$, there is sufficient empirical information to determine a unique parameter estimate, and the system should be purged of the pseudo information.

In one such a method, the dispersion matrix P_t of the estimated parameter vector is resolved into two components such that $P_t = P_t^* + \rho P_t^\circ$, where P_t^* relates to the sample information and P_t° relates to the diffuse presample information. The latter is used to initialise the recursive process at time $t = 0$. As observations accrue, the new information is incorporated into P_t^* and any conflicting pseudo information is removed from P_t°.

To implement the updating formulae, we need expressions for f_t^{-1} and κ_t that reflect the nature of the information. Let

$$f_t = f_t^* + \rho f_t^\circ \qquad \text{with} \qquad f_t^* = x_t'P_{t-1}^*x_t + 1, \qquad f_t^\circ = x_t'P_{t-1}^\circ x_t. \tag{22}$$

On the assumption that $f_t^\circ \neq 0$, there is $f_t = \rho f_t^\circ(1 - \rho^{-1}q)$ with $q = -f_t^*/f_t^\circ$. Since $\rho > 1$, there is a series expansion of the inverse of the form

$$f_t^{-1} = \frac{1}{\rho f_t^\circ}\left(1 + \frac{q}{\rho} + \frac{q^2}{\rho^2} + \cdots\right) \tag{23}$$

$$= \frac{g_1}{\rho} + \frac{g_2}{\rho^2} + \frac{g_3}{\rho^3} + \cdots.$$

To find the terms of this expansion, consider the equation $1 = f_t f_t^{-1}$ written as

$$\begin{aligned} 1 &= (f_t^* + \rho f_t^\circ)\left(\frac{g_1}{\rho} + \frac{g_2}{\rho^2} + \frac{g_3}{\rho^3} + \cdots\right) \\ &= f_t^\circ g_1 + \frac{1}{\rho}(f_t^* g_1 + f_t^\circ g_2) + \frac{1}{\rho^2}(f_t^* g_2 + f_t^\circ g_3) + \cdots. \end{aligned} \tag{24}$$

Here, the first term in the product on the RHS is unity, whereas the remaining terms, associated with negative powers of ρ, are zeros. It follows that

$$g_1 = (f_t^\circ)^{-1} \quad \text{and} \quad g_2 = -(f_t^\circ)^{-2} f_t^*. \tag{25}$$

One can ignore g_3 and the coefficients associated with higher powers of $1/\rho$, which vanish from all subsequent expressions as ρ increases. Next, there is

$$\begin{aligned} \kappa_t &= P_{t-1} x_t f_t^{-1} = (P_{t-1}^* x_t + \rho P_{t-1}^\circ x_t)\left(\frac{g_1}{\rho} + \frac{g_2}{\rho^2} + \frac{g_3}{\rho^3} + \cdots\right) \\ &= P_{t-1}^\circ x_t g_1 + \frac{1}{\rho}(P_{t-1}^* g_1 + P_{t-1}^\circ g_2) x_t + \frac{1}{\rho^2}(P_{t-1}^* g_2 + P_{t-1}^\circ g_3) x_t + \cdots \\ &= d_0 + \frac{d_1}{\rho} + \frac{d_2}{\rho^2} + \cdots, \end{aligned} \tag{26}$$

where

$$d_0 = P_{t-1}^\circ x_t (f_t^\circ)^{-1} \quad \text{and} \quad d_1 = P_{t-1}^* x_t (f_t^\circ)^{-1} - P_{t-1}^\circ x_t (f_t^\circ)^{-2} f_t^*. \tag{27}$$

As $\rho \to \infty$, only the first term of (26) survives, giving $\kappa_t = P_{t-1}^\circ x_t (f_t^\circ)^{-1} = \kappa_t^\circ$. Therefore, when $f_t^\circ \neq 0$, the updating equation for the parameter estimate is

$$b_t = b_{t-1} + P_{t-1}^\circ x_t (f_t^\circ)^{-1} h_t. \tag{28}$$

Finally, consider the updating equation for the dispersion of the estimate. This embodies

$$\begin{aligned} \kappa_t x_t' P_{t-1} &= \left(d_0 + \frac{d_1}{\rho} + \frac{d_2}{\rho^2} + \cdots\right)(x_t' P_{t-1}^* + \rho x_t' P_{t-1}^\circ) \\ &= \rho d_0 x_t' P_{t-1}^\circ + (d_0 x_t' P_{t-1}^* + d_1 x_t' P_{t-1}^\circ) + \cdots. \end{aligned} \tag{29}$$

Putting the leading terms into (14) and separating $P_t = P_t^* + \rho P_t^\circ$ into its two parts gives

$$P_t^\circ = P_{t-1}^\circ - P_{t-1}^\circ x_t (f_t^\circ)^{-1} x_t' P_{t-1}^\circ, \tag{30}$$

$$\begin{aligned} P_t^* = P_{t-1}^* &+ P_{t-1}^\circ x_t (f_t^\circ)^{-1} f_t^* (f_t^\circ)^{-1} x_t' P_{t-1}^\circ \\ &- P_{t-1}^\circ x_t (f_t^\circ)^{-1} x_t' P_{t-1}^* - P_{t-1}^* x_t (f_t^\circ)^{-1} x_t' P_{t-1}^\circ. \end{aligned} \tag{31}$$

Using the notation

$$\kappa_t^\circ = P_{t-1}^\circ x_t (x_t' P_{t-1}^\circ x_t)^{-1} \qquad \text{and} \qquad \Lambda_t^\circ = I - \kappa_t^\circ x_t', \tag{32}$$

equation (28) may be written as

$$\begin{aligned} b_t &= b_{t-1} + \kappa_t^\circ (y_t - x_t' b_{t-1}) \\ &= \Lambda_t^\circ b_{t-1} + \kappa_t^\circ y_t. \end{aligned} \tag{33}$$

Using the same notation, equations (30) and (31) can be written as

$$P_t^\circ = \Lambda_t^\circ P_{t-1}^\circ, \tag{34}$$

$$P_t^* = \Lambda_t^\circ P_{t-1}^* \Lambda_t^{\circ\prime} + \kappa_t^\circ \kappa_t^{\circ\prime}. \tag{35}$$

The updating equation of (34), which is associated with the diffuse information, has the form of $P_t^\circ = (I - \kappa_t^\circ x_t') P_{t-1}^\circ$, where $\kappa_t^\circ x_t' = P_{t-1}^\circ x_t (x_t' P_{t-1}^\circ x_t)^{-1} x_t'$ and $I - \kappa_t^\circ x_t'$ are idempotent matrices. Thus, P_t° is formed by projecting P_{t-1}° onto the subspace orthogonal to x_t, with the result that

$$x_s' P_t^\circ = 0 \qquad \text{when} \qquad t \geq s. \tag{36}$$

Unless $\kappa_t^\circ x_t' = 0$ the matrix $\Lambda_t^\circ = I - \kappa_t^\circ x_t'$ will have less than full rank. If the vectors $x_t'; t = 1, \ldots, k$ are linearly independent, then after k steps, the loss of rank will lead to

$$\prod_{j=1}^{k} \Lambda_j^\circ = \prod_{j=1}^{k} (I - \kappa_j^\circ x_j') = 0, \tag{37}$$

and, therefore, to $P_k^\circ = \prod_{j=1}^{k} (I - \kappa_j^\circ x_j') P_0^\circ = 0$. Beyond that point, there will be $f_t^\circ = x_t' P_{t-1}^\circ x_t = 0$ and, therefore, $f_t = f_t^*$. It follows from the logic of the preceding derivation that the recursive equations will assume the standard forms of (7) and (9).

In the absence informative prior information, the procedure can be initialised with $P_0^* = 0$, $P_0^\circ = I$ and with an arbitrary value for b_0. With these initialisations, the algorithm gives $b_k = X_k^{-1} Y_K$ and $P_k = (X_k' X_k)^{-1}$ when $t = k$, regardless of the starting value b_0.

The algorithm that we have described was proposed in Ansley and Kohn (1985a), where it was developed in the context of the Kalman filter. Koopman (1997) provides some elaborations, and an accessible exposition is in Durbin and Koopman (2001).

Example. To illustrate the process of initialisation, consider the case of $k = 3$.

Running the recursion (33) for three iterations, we get

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \Lambda_1^\circ \\ \Lambda_2^\circ\Lambda_1^\circ \\ \Lambda_3^\circ\Lambda_2^\circ\Lambda_1^\circ \end{bmatrix} b_0 + \begin{bmatrix} \kappa_1^\circ & 0 & 0 \\ \Lambda_2^\circ\kappa_1^\circ & \kappa_2^\circ & 0 \\ \Lambda_3^\circ\Lambda_2^\circ\kappa_1^\circ & \Lambda_3^\circ\kappa_2^\circ & \kappa_3^\circ \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{38}$$

To prove that that $b_3 = X_3^{-1}Y_3$, regardless of the value of b_0, we must show that

$$\text{(i)} \quad \Lambda_3^\circ\Lambda_2^\circ\Lambda_1^\circ = 0 \qquad \text{and} \qquad \text{(ii)} \quad \begin{bmatrix} \Lambda_3^\circ\Lambda_2^\circ\kappa_1^\circ & \Lambda_3^\circ\kappa_2^\circ & \kappa_3^\circ \end{bmatrix} = X_3^{-1}. \tag{39}$$

Here, (i) is subsumed under (37). To demonstrate (ii), consider the fact that, in view of (36),

$$\kappa_t^\circ = P_{t-1}^\circ x_t (f_t^\circ)^{-1} \qquad \text{has} \qquad x_s'\kappa_t^\circ = \begin{cases} 1, & \text{if } t = s, \\ 0, & \text{if } t > s, \end{cases} \tag{40}$$

and , consequently,

$$\Lambda_t^\circ = I - \kappa_t^\circ x_t' \qquad \text{has} \qquad x_s'\Lambda_t^\circ = \begin{cases} 0, & \text{if } t = s, \\ x_s', & \text{if } t > s. \end{cases} \tag{41}$$

It follows immediately that

$$X_3X_3^{-1} = \begin{bmatrix} x_1'\Lambda_3^\circ\Lambda_2^\circ\kappa_1^\circ & x_1'\Lambda_3^\circ\kappa_2^\circ & x_1'\kappa_3^\circ \\ x_2'\Lambda_3^\circ\Lambda_2^\circ\kappa_1^\circ & x_2'\Lambda_3^\circ\kappa_2^\circ & x_2'\kappa_3^\circ \\ x_3'\Lambda_3^\circ\Lambda_2^\circ\kappa_1^\circ & x_3'\Lambda_3^\circ\kappa_2^\circ & x_3'\kappa_3^\circ \end{bmatrix} = I, \tag{42}$$

which proves (ii).

Next, we wish to show that that $P_3 = P_3^* = (X_3'X_3)^{-1}$. Consider the first three iterations of (35). With $P_0^\circ = I$ and $P_0^* = 0$, we get

$$\begin{aligned} P_1^* &= \kappa_1^\circ\kappa_1^{\circ\prime}, \\ P_2^* &= \Lambda_2^\circ\kappa_1^\circ\kappa_1^{\circ\prime}\Lambda_2^{\circ\prime} + \kappa_2^\circ\kappa_2^{\circ\prime}, \\ P_3^* &= \Lambda_3^\circ\Lambda_2^\circ\kappa_1^\circ\kappa_1^{\circ\prime}\Lambda_2^{\circ\prime}\Lambda_3^{\circ\prime} + \Lambda_3^\circ\kappa_2^\circ\kappa_2^{\circ\prime}\Lambda_3^{\circ\prime} + \kappa_3^\circ\kappa_3^{\circ\prime}. \end{aligned} \tag{43}$$

Reference to (39.ii) shows that the last of these is just the product $X_3^{-1}(X_3^{-1})'$, which is the result that we are seeking.

5 The Prediction-Error Decomposition

The equations of the regression model containing the full set of observations can be written in the familiar form of $y = X\beta + \varepsilon$, where $E(\varepsilon) = 0$

and $D(\varepsilon) = \sigma^2 I$. When a prior distribution is available for β, we also have $E(\beta) = b_0$ and $D(\beta) = \sigma^2 P_0$. Combining these gives

$$\begin{aligned} E(y) &= XE(\beta) + E(\varepsilon) \\ &= Xb_0 \end{aligned} \quad \text{and} \quad \begin{aligned} D(y) &= XD(\beta)X' + D(\varepsilon) \\ &= \sigma^2 XP_0X' + \sigma^2 I. \end{aligned} \tag{44}$$

Assuming that the stochastic elements are normally distributed, the marginal density function of y is

$$N(y) = (2\pi\sigma)^{-T/2}|XP_0X' + I|^{-1/2} \exp\{-S(y)/(2\sigma^2)\}, \tag{45}$$

whose quadratic exponent is

$$\begin{aligned} S(y) &= (y - Xb_0)'(XP_0X' + I)^{-1}(y - Xb_0) \\ &= (y - Xb_0)'\{I - X(X'X + P_0^{-1})^{-1}X'\}(y - Xb_0). \end{aligned} \tag{46}$$

The second equality follows from (A.3.iii).

The recursive regression algorithm (10)–(14) entails a decomposition of the marginal function $N(y)$ called the prediction-error decomposition. This takes the form

$$N(y_1, \ldots, y_T; \mathcal{I}_0) = N(y_1; \mathcal{I}_0) \prod_{t=2}^{T} N(y_t|\mathcal{I}_{t-1}). \tag{47}$$

For $t > 1$, the factors on the RHS take the form

$$N(y_t|\mathcal{I}_{t-1}) = (2\pi\sigma^2 f_t)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\frac{(y_t - x_t'b_{t-1})^2}{1 + x_t'P_{t-1}x_t}\right\}. \tag{48}$$

The marginal density function $N(y_1; \mathcal{I}_0)$, which is the first factor of the decomposition, is obtained by specialising (45) to the case of a single observation or, equally, by setting $t = 1$ in $N(y_t|\mathcal{I}_{t-1})$. Thus, the quadratic function in (46) can be written alternatively as

$$S(y) = \sum_{t=1}^{T} \frac{(y_t - x_t'b_{t-1})^2}{1 + x_t'P_{t-1}x_t} = \sum_{t=1}^{T} \frac{h_t^2}{f_t} = \sum_{t=1}^{T} w_t^2. \tag{49}$$

A one-to-one correspondence can be demonstrated between the errors $y_t - x_t'b_0$ and the prediction errors $h_t = y_t - x_t'b_{t-1}$. Consider recursive formula

$$\begin{aligned} b_t &= b_{t-1} + \kappa_t(y_t - x_t'b_{t-1}) \\ &= \Lambda_t b_{t-1} + \kappa_t y_t, \end{aligned} \tag{50}$$

where $\Lambda_t = I - \kappa_t x_t'$. Running the recursion for the first few iterations, we get

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \Lambda_1 \\ \Lambda_2\Lambda_1 \\ \Lambda_3\Lambda_2\Lambda_1 \end{bmatrix} b_0 + \begin{bmatrix} \kappa_1 & 0 & 0 \\ \Lambda_2\kappa_1 & \kappa_2 & 0 \\ \Lambda_3\Lambda_2\kappa_1 & \Lambda_3\kappa_2 & \kappa_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{51}$$

Then, since $h_t = y_t - x_t'b_{t-1}$,

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -x_2'\kappa_1 & 1 & 0 & 0 \\ -x_3'\Lambda_2\kappa_1 & -x_3'\kappa_2 & 1 & 0 \\ -x_4'\Lambda_3\Lambda_2\kappa_1 & -x_4'\Lambda_3\kappa_2 & -x_4'\kappa_3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} - \begin{bmatrix} x_1' \\ x_2'\Lambda_1 \\ x_3'\Lambda_2\Lambda_1 \\ x_4'\Lambda_3\Lambda_2\Lambda_1 \end{bmatrix} b_0. \tag{52}$$

On defining $\Lambda_{j,m} = \Lambda_j\Lambda_{j-1}\cdots\Lambda_m$, with $\Lambda_{j,j} = \Lambda_j$ and $\Lambda_{j,j+1} = I$, the generic expression for the prediction error becomes

$$\begin{aligned} h_t &= y_t - x_t'b_{t-1} \\ &= y_t - x_t'\Lambda_{t-1,1}b_0 - x_t'\sum_{j=1}^{t-1}\Lambda_{t-1,j+1}\kappa_j y_j. \end{aligned} \tag{53}$$

Equation (52) can be summarised as $h = Ly - Wb_0$. But $E(h) = 0$ and $E(y) = Xb_0$, so the equation indicates that $LXb_0 = Wb_0$, or $W = LX$, since b_0 can take any value. (The equality $W = LX$ can also be demonstrated algebraically without resort to the expectations operator.) Substituting this back into the original equation gives $h = L(y - Xb_0)$, which holds for any extension of the recursion. This establishes the relationship between the errors $y_t - x_t'b_0$ and the prediction errors $h_t = y_t - x_t'b_{t-1}$. Thus, the marginal sum of squares of (46) can also be written as

$$\begin{aligned} S(y) &= (y - Xb_0)'(XP_0X' + I)^{-1}(y - Xb_0) \\ &= (y - Xb_0)'L'F^{-1}L(y - Xb_0) = h'F^{-1}h, \end{aligned} \tag{54}$$

where $\sigma^2 F = \sigma^2\text{diag}\{f_1, \ldots, f_T\}$ is the matrix of the prediction-error dispersions.

The case with no prior information on β may be handled by concentrating the likelihood function $N(y)$ in respect of b_0 and P_0. The minimising value for b_0 is the ordinary least-squares estimator $b = (X'X)^{-1}X'y$, as will be demonstrated in Section 8, and the minimising value for P_0 is zero.

The condition $P_0 = 0$ normally signifies that there is complete information regarding β. This is clearly at variance with the actual circumstance of no prior information. This anomaly indicates that the appropriate way to estimate β in the absence of prior information is by minimising the conditional function $S(y|\beta) = (y - X\beta)'(y - X\beta)$ instead of the marginal function $S(y)$.

Setting $\beta = b_0 = b$ reduces both $S(y)$ and $S(y|\beta)$ to the concentrated function

$$S^c(y) = e'e = y'\{I - X(X'X)^{-1}X'\}y = \varepsilon'\{I - X(X'X)^{-1}X'\}\varepsilon, \tag{55}$$

where $e = [e_1, \ldots, e_T]'$ stands for the vector of ordinary least-squares residuals.

In the absence of prior information, the concentrated function has a prediction-error decomposition of the form (49), but the index of summation begins at $t = k + 1$, instead of $t = 1$, and the starting values are $b_k = X_k^{-1}Y_k$ and $P_k = (X_k'X_k)^{-1}$—see Pollock (1999, p. 231). The notation $X = [X_1', X_2']'$, $y = [y_1', y_2']'$, where $X_1' = [x_1, \ldots, x_k]'$ and $y_1' = [y_1, \ldots, y_k]'$, may be used to denote the partition of the sample into the first k elements and the remainder. Then, the starting values become $b_1 = X_1^{-1}y_1$ and $P_1 = (X_1'X_1)^{-1}$, and an expression for $S^c(y)$ arises that is analogous to that of (54):

$$\begin{aligned} S^c(y) &= (y_2 - X_2b_1)'\{X_2(X_1'X_1)^{-1}X_2' + I\}^{-1}(y_2 - X_2b_1) \\ &= (y_2 - X_2b_1)'L_2'F_2^{-1}L_2(y_2 - X_2b_1) = h_2'F_2^{-1}h_2. \end{aligned} \tag{56}$$

Here, L_2 and $F_2 = \text{diag}\{f_{k+1}, \ldots, f_T\}$ are analogous to the matrices defined in respect of (54). The vector $h_2 = [h_{k+1}, \ldots, h_T]'$ contains the prediction errors, whose normalised versions $w_t = h_t/f_t$ are in the vector w.

In the absence of prior information concerning the regression parameters, the normalised prediction errors are conventionally described as *recursive residuals.* The essential conditions affecting the recursive residuals are that

$$E(w) = 0 \quad \text{and} \quad D(w) = \sigma^2 I_{T-k}, \tag{57}$$

which is to say that they possess a spherical distribution.

There are various alternative residuals associated with the classical regression model that have statistical properties similar to those of the recursive residuals and which can also be used for testing the assumptions of the model. Thus, Theil (1971) has defined the LUS class of linear unbiased residuals with a scalar covariance matrix (i.e. a scalar multiple of the identity matrix). It is helpful, for later reference, to demonstrate how these are derived.

Observe that, since $X'X$ is a full-rank symmetric matrix of order k, there exists a matrix T such that $T'T = (X'X)^{-1}$. Therefore, $X(X'X)^{-1}X' = XT'TX' = C_1C_1'$, where C_1 is a $T \times k$ matrix of orthonormal vectors such that $C_1'C_1 = I_k$. Let C_2 be the $T \times (T - k)$ orthogonal complement to C_1 so that $C_2'C_1 = C_2'X = 0$, $C_2'C_2 = I_{T-k}$ and $C_1C_1' + C_2C_2' = I_T$. Then,

$$\begin{aligned} S^c(y) &= \sum_{t=1}^{T} e_t^2 = y'\{I - X(X'X)^{-1}X'\}y \\ &= y'C_2C_2'y = \sum_{t=k+1}^{T} v_t^2. \end{aligned} \tag{58}$$

This equation relates the ordinary least-squares residuals $C_2C_2'y = e = [e_1, \ldots, e_T]'$, to the LUS residuals $C_2'y = C_2'e = v = [v_{k+1}, \ldots, v_T]'$.

Now observe that $v = C_2'(y - X\beta) = C_2'\varepsilon$. Since $E(\varepsilon\varepsilon') = \sigma^2 I_T$ and $C_2'C_2 = I_{T-k}$, it follows that

$$E(v) = 0 \quad \text{and} \quad D(v) = C_2'E(\varepsilon\varepsilon')C_2 = \sigma^2 I_{T-k}, \tag{59}$$

which shows that the LUS residuals possess a spherical distribution. Indeed, the recursive residuals are just an instance of the LUS residuals. An explicit expression for the matrix C in this case has been given by Dufour (1982).

Since they are independently and identically distributed under the assumptions of the regression model, the recursive residuals enable exact tests of the assumptions to be derived with ease. Harvey (1990) indicates that the recursive residuals are amenable to an exact von Neumann ratio test aimed at detecting serial correlation in the disturbances, which is preferable to the Durbin–Watson test constructed from the ordinary least-squares residuals. Since the least-squares residuals are dependent on the values in X, it is not possible to derive exact significance points that apply to every instance of that test.

Another leading use of recursive residuals is in the CUSUM test, proposed by Brown, Durbin and Evans (1975). This detects instability in regression parameters, rejecting the hypothesis of invariance if the trajectory of the cumulative sum of the recursive residuals crosses an upper or lower critical line. The lines are calculated with reference to the boundary-crossing probabilities of a Brownian motion defined on a unit interval, which approximates the CUSUM process with increasing accuracy as the sample size increases—see Durbin (1971).

A simple alternative to the CUSUM statistic is provided by the ratio

$$t = \frac{\sum_{t=k+1}^{T} w_t/\sqrt{T-k}}{\left\{\sum_{t=k+1}^{T}(w_t - \bar{w})^2/(T-k-1)\right\}^{1/2}}, \tag{60}$$

where $\bar{w}$ is the arithmetic mean of the recursive residuals. This statistic, proposed in Harvey and Collier (1977), is distributed as Students t with $T-k-1$ degrees of freedom under the null of parametric constancy.

The use of recursive residuals for detecting functional misspecification and parametric change has been further investigated in Dufour (1982) and Krämer, Ploberger and Alt (1988). The latter assesses the use of the CUSUM test when there are lagged dependent variables among the regressors, and shows that the test retains its asymptotic significance levels in dynamic models.

A closely related test is the fluctuations test of Ploberger, Krämer and Kontros (1989), which is based on successive parameter estimates rather than on recursive residuals. It can be seen, in reference to (16) and (18), that the differences between successive parameter estimates, which are elements of the

vectors $b_t - b_{t-1} = \kappa_t(y_t - x_t'b_{t-1})$, are scalar functions of the recursive residuals. Dufour (1982) has recommended that one should track the trajectories of these elements.

The fluctuations test is based upon the deviations of the current estimates from the full-sample estimate. These quantities $b_t - b_T = \sum_{s=t+1}^{T}(b_{s-1} - b_s)$, bear a one-to-one relationship with the vectors of differences. The test is based on the maximum value of the deviations. Developments and extensions of the test have been provided in Kuan and Hornik (1995) and in Kuan (1998).

The techniques of recursive estimation are exploited in Banerjee, Lumsdaine and Stock (1992) in specification tests for nonstationary dynamic models. Their aim is to determine whether the data are best described by a trend-stationary model or a difference-stationary model with a unit root within an autoregressive operator, which is their null hypothesis. (The models of the null and alternative hypotheses are nested within a comprehensive model in the manner of Bhargava (1986).) They also devise tests to investigate the possibility that the time series is stationary around a broken trend line.

The test of the unit-root hypothesis entails a recursive calculation of the Dickey–Fuller (1979) statistics. The trajectories of the test statistics under the null hypothesis are depicted in terms of Brownian motion on the unit interval. In this respect, the work adopts the methodology of Brown, Durbin and Evans (1975). The applicability of such an approach to the theory of time-series models with autoregressive unit roots is demonstrated in Phillips (1987), which has become the mainstay of many subsequent econometric studies of integrated and co-integrated time series. An extensive survey of the econometric literature on unit roots, structural breaks and trends has been provided in Stock (1994).

6 Extensions of the Recursive Least-Squares Algorithm

The algorithm presented in the previous sections represents little more than an alternative means for computing the ordinary least-squares regression estimates. If the parameters of the process generating the data are constant, then we can expect the estimate b_t to converge to a limit as the number of observations t increases. At the same time, the elements of the dispersion matrix $\sigma^2 P_t$ will decrease in value, as will the filter gain κ_t. Thus, the impact of successive prediction errors on the estimate of β diminishes as more information is included.

If there is doubt about the constancy of the regression parameter, it may appropriate to discard data that have reached a date of expiry. 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 or moving-window regression. Implementations are available in recent versions of the more popular econometric computer packages such as *Microfit 4.0* and *PCGive 10.0*.

To extend the algorithm of the previous section to produce a rolling regression, one needs to remove the data that were acquired at time $t-n$. The first step is to adjust the moment matrix to give $P_t^{*-1} = P_{t-1}^{-1} - x_{t-n}x'_{t-n}$. The matrix inversion formula of (A.3.ii) indicates that

$$\begin{aligned} P_t^* &= (P_{t-1}^{-1} - x_{t-n}x'_{t-n})^{-1} \\ &= P_{t-1} + P_{t-1}x_{t-n}(x'_{t-n}P_{t-1}x_{t-n} - 1)^{-1}x'_{t-n}P_{t-1}. \end{aligned} \tag{61}$$

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

$$\begin{aligned} b_t^* &= b_{t-1} - P_t^*x_{t-n}(y_{t-n} - x'_{t-n}b_{t-1}) \\ &= b_{t-1} - P_{t-1}x_{t-n}(x'_{t-n}P_{t-1}x_{t-n} - 1)^{-1}(y_{t-n} - x'_{t-n}b_{t-1}). \end{aligned} \tag{62}$$

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 first expression for b_t^* on the RHS of (62). The second expression depends on the identity $(P_{t-1}^{-1} - x_{t-n}x'_{t-n})^{-1}x_{t-n} = P_{t-1}x_{t-n}(x'_{t-n}P_{t-1}x_{t-n} - 1)^{-1}$, which is in the form of $a^{-1}c = bd^{-1}$ and which can be confirmed by recasting it as $cd = ab$. Finally, the estimate b_t, which is based on the n data points $x_t, \ldots, x_{t-n+1}$, is obtained from (7) by replacing b_{t-1} with b_t^* and P_{t-1} with P_t^*.

The method of rolling regression is useful for initialising an ordinary recursive regression that lacks prior information for the regression parameters. A rolling regression can be set in motion using pseudo information, such as $b_0 = 0$ and $P_0 = I$. Then, as the regression rolls forwards, the pseudo information is replaced by sample information until $t = k$, at which point there is only sample information in the data window. Then, the rolling regression can be converted to an ordinary recursive regression with $b_k = X_k^{-1}Y_k$ and $P_k = (X'_kX_k)^{-1}$. This use of the rolling regression algorithm, which is a straightforward extension of the recursive algorithm, allows one to dispense with a matrix inversion routine in finding the initial values.

In econometrics, increasing use is being made of test statistics based upon rolling regression. Banerjee, Lumsdaine and Stock (1992), for example, have accompanied recursive tests with ones based on rolling regressions. However, they are not explicit about the exact nature of the alternative hypotheses that motivate such tests.

This matter is elucidated in Chu, Hornik and Kuan (1995), where the possibility of a temporary parameter shift within a regression model with stationary explanatory variables is considered. Such shifts can be overlooked by recursive statistics if their values are too strongly influenced by a stable past. They are more likely to be detected via the fluctuations of moving estimates computed from a sequence of subsamples demarcated by a rolling data window. Under the null hypothesis of parametric constancy, the deviations of the rolling estimates from the full-sample estimates converge weakly in prob-

ability to the increments of a Brownian bridge. This provides the basis for determining the critical values of the tests.

Smith and Taylor (2001) uses the techniques of recursive and rolling regression in testing the constancy of a seasonal process described in terms of an autoregressive model with complex roots whose arguments correspond to the seasonal frequency and its harmonics. The paper describes tests of the hypothesis that the seasonal process entails roots of unit modulus against an alternative of stable roots for part of its history in respect of some, if not all, of the seasonal frequencies.

The test statistics are modelled on those in Hylleberg, Engle, Granger and Yoo (1990), which describes a structure within which hypotheses relating to the various seasonal roots may be tested individually, via t-tests in the case of real-valued roots, or via F-tests in case of conjugate complex roots. (Alternative test statistics, that entertain the null hypothesis of no seasonal unit roots, are proposed in Canova and Hansen (1995) which adapts the statistics of Kwiatkowski, Phillips, Schmidt and Shin (1992) that are aimed at detecting real-valued roots.)

The tests of Smith and Taylor are based on maximum and minimum values from sequences of t and F-statistics generated by recursive and rolling regressions, running in both directions, together with the differences of these values. Their approach to deriving the critical values for their tests is via Brownian motion described on the unit interval. This is a modern alternative to the methods used in Dickey, Hasza and Fuller (1984) for developing tests of the null hypothesis that there are unit roots at every seasonal frequency against an alternative hypothesis of no seasonal unit roots.

Discarding observations beyond a date of expiry is appropriate when the processes generating the data are liable to undergo sudden structural changes. It ensures that any misinformation conveyed by the data that predates 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 $\lambda \in (0, 1]$ be the factor by which the data are discounted from one period to the next. Then the expression for P_t in (9) would be replaced by

$$\begin{aligned} P_t &= (\lambda P_{t-1}^{-1} + x_t x_t')^{-1} \\ &= \frac{1}{\lambda}\Big\{P_{t-1} - P_{t-1}x_t(x_t'P_{t-1}x_t + \lambda)^{-1}x_t'P_{t-1}\Big\}. \end{aligned} \tag{63}$$

The formula for the parameter estimate becomes

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + \lambda)^{-1}(y - x_t'b_{t-1}). \tag{64}$$

Discounted regression has yet to achieve widespread use in econometrics. It has been used extensively in adaptive control, beginning with Åström, Borisson, Ljung and Wittenmark (1977). Its purpose in this context is to prevent

the recursive estimator from converging and to accommodate parametric drift in the system subject to control. Examples are provided in Kiparissides and Shah (1983) and Wellstead and Zarrop (1991).

Lozano (1983) provides an analysis of the convergence of discounted least squares under favourable conditions of persistent excitation. This shows the dispersion of the estimated regression parameters tending to constancy. However, a problem arises with a constant forgetting factor if the system is parametrically stable and the inputs become quiescent. Then the old information is forgotten while little new information is added. This can make the control system overly sensitive to disturbances and susceptible to numerical and computational difficulties, symptomised by an explosive growth in elements of the dispersion matrix of the regression estimate.

The problem can be solved by devising systems of variable forgetting factors aimed at maintaining a constant information content within successive estimates. Such systems are analysed in Zarrop (1983), Sanoff and Wellstead (1983) and Canetti and España (1989); and Fortescue, Kershenbaum and Ydstie (1981) describe an implementation. More sophisticated memory shaping systems are possible that will allow the information content to grow indefinitely if there is no hint of parametric inconstancy and that discard information rapidly when there is clear evidence of change.

A belief in the parametric constancy of economic systems might not be the only reason why econometricians have proved resistant to devices such as discounted regression. Whereas occasional structural breaks can be accommodated easily, continuous structural change is liable to subvert the very objectives of structural econometric analysis. Also, both rolling regression and discounted regression are incapable of producing estimates that are statistically consistent, although, as noted, this objection may be overcome by sophisticated memory shaping.

A final objection to the algorithms of recursive regression concerns their laggardly and backward-looking nature. Recursive regressions that hold only past data in their memories are liable to react to structural changes with considerable delay. This objection can be overcome if one is prepared to look forward in time as well as backward by replacing recursive regression by a combination of the Kalman filter, which is backward-looking, and its associated smoothing algorithms, which are forward-looking.

7 The Kalman Filter

The basic equations of the Kalman filter will be derived in the briefest possible manner. The state-space model that underlies the Kalman filter consists of two equations

$$y_t = H_t\beta_t + \eta_t, \qquad \textit{Observation Equation} \tag{65}$$

$$\beta_t = \Phi_t\beta_{t-1} + \nu_t, \qquad \textit{Transition Equation} \tag{66}$$

where y_t is a vector of observations on the system and β_t is the state vector of k elements. The observation error η_t and the state disturbance ν_t are mutually uncorrelated, normally distributed, random vectors of zero mean with dispersion matrices

$$D(\eta_t) = \Omega_t \qquad \text{and} \qquad D(\nu_t) = \Psi_t. \tag{67}$$

The observation equation is analogous to the regression equation of (1), but y_t may be a vector quantity. The transition equation is new.

It is assumed that the matrices H_t, Φ_t, Ω_t and Ψ_t are known for all $t = 1, \ldots, T$ and that an initial estimate $E(\beta_0) = b_0$ is available for the state vector β_0 at $t = 0$ together with a dispersion matrix $D(\beta_0) = P_0$. The initial information is $\mathcal{I}_0$. The information available at time t is $\mathcal{I}_t = \{y_t, \ldots, y_1, \mathcal{I}_0\}$.

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

$$b_{t|t-1} = \Phi_t b_{t-1}, \qquad \textit{State Prediction} \tag{68}$$

$$P_{t|t-1} = \Phi_t P_{t-1} \Phi_t' + \Psi_t, \qquad \textit{Prediction Dispersion} \tag{69}$$

$$e_t = y_t - H_t b_{t|t-1}, \qquad \textit{Prediction Error} \tag{70}$$

$$F_t = H_t P_{t|t-1} H_t' + \Omega_t, \qquad \textit{Error Dispersion} \tag{71}$$

$$K_t = P_{t|t-1} H_t' F_t^{-1}, \qquad \textit{Kalman Gain} \tag{72}$$

$$b_t = b_{t|t-1} + K_t e_t, \qquad \textit{State Estimate} \tag{73}$$

$$P_t = (I - K_t H_t) P_{t|t-1}. \qquad \textit{Estimate Dispersion} \tag{74}$$

It is useful to define

$$\Lambda_t = (I - K_t H_t)\Phi_t. \tag{75}$$

There are two additions to the recursive regression algorithm (10)–(14): equation (68) for the state prediction and equation (69) for its dispersion. These arise from the transition equation (66); and they vanish if $\Phi = I$, $\nu_t = 0$ and $D(\nu_t) = \Psi_t = 0$ so that $P_{t|t-1}$ becomes P_{t-1} in the remaining equations.

The Kalman filter can be derived using the algebra of conditional expectations, given in (A.8). Amongst (68)–(74), equations (70) and (72) are merely definitions. To demonstrate (68), use (A.8.iii) to show that

$$\begin{aligned} E(\beta_t|\mathcal{I}_{t-1}) &= E\{E(\beta_t|\beta_{t-1})|\mathcal{I}_{t-1}\} \\ &= E\{\Phi_t \beta_{t-1}|\mathcal{I}_{t-1}\} \\ &= \Phi_t b_{t-1}. \end{aligned} \tag{76}$$

Use (A.8.v) to demonstrate (69):

$$
\begin{aligned}
D(\beta_t|\mathcal{I}_{t-1}) &= D(\beta_t|\beta_{t-1}) + D\{E(\beta_t|\beta_{t-1})|\mathcal{I}_{t-1}\} \\
&= \Psi_t + D\{\Phi_t\beta_{t-1}|\mathcal{I}_{t-1}\} \\
&= \Psi_t + \Phi_t P_{t-1}\Phi_t'.
\end{aligned} \tag{77}
$$

To obtain (71), substitute (65) into (70) to give $e_t = H_t(\beta_t - b_{t|t-1}) + \eta_t$. Then, in view of the statistical independence of the terms on the RHS, one has

$$
\begin{aligned}
D(e_t) &= D\{H_t(\beta_t - b_{t|t-1})\} + D(\eta_t) \\
&= H_t P_{t|t-1} H_t' + \Omega_t = D(y_t|\mathcal{I}_{t-1}).
\end{aligned} \tag{78}
$$

To demonstrate the updating equation (73), begin by noting that

$$
\begin{aligned}
C(\beta_t, y_t|\mathcal{I}_{t-1}) &= E\{(\beta_t - b_{t|t-1})y_t'\} \\
&= E\{(\beta_t - b_{t|t-1})(H_t\beta_t + \eta_t)'\} \\
&= P_{t|t-1}H_t'.
\end{aligned} \tag{79}
$$

It follows from (A.8.i) that

$$
\begin{aligned}
E(\beta_t|\mathcal{I}_t) &= E(\beta_t|\mathcal{I}_{t-1}) + C(\beta_t, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\{y_t - E(y_t|\mathcal{I}_{t-1})\} \\
&= b_{t|t-1} + P_{t|t-1}H_t'F_t^{-1}e_t.
\end{aligned} \tag{80}
$$

The dispersion matrix in (74) for the updated estimate is obtained via (A.8.ii):

$$
\begin{aligned}
D(\beta_t|\mathcal{I}_t) &= D(\beta_t|\mathcal{I}_{t-1}) - C(\beta_t, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})C(y_t, \beta_t|\mathcal{I}_{t-1}) \\
&= P_{t|t-1} - P_{t|t-1}H_t'F_t^{-1}H_tP_{t|t-1}.
\end{aligned} \tag{81}
$$

It is useful for later analysis to express the current state vector in terms of the initial state vector and a sequence of state disturbances. By repeated back substitution in (66), we obtain

$$
\beta_t = \Phi_{t,1}\beta_0 + \sum_{j=1}^{t} \Phi_{t,j+1}\nu_j, \tag{82}
$$

where $\Phi_{t,j+1} = \Phi_t \cdots \Phi_{j+1}$, with $\Phi_{j,j} = \Phi_j$ and $\Phi_{j,j+1} = I$. Substituting this into the equation $y_t = H_t\beta_t + \eta_t$ from (65) gives another useful expression:

$$
\begin{aligned}
y_t &= H_t\Phi_{t,1}\beta_0 + H_t\sum_{j=1}^{t}\Phi_{t,j+1}\nu_j + \eta_t \\
&= X_t\beta_0 + \varepsilon_t.
\end{aligned} \tag{83}
$$

On defining the vectors $y = [y_1', \ldots, y_T']'$, $\varepsilon = [\varepsilon_1', \ldots, y_T']'$ and the matrix $X = [X_1', \ldots, X_T']'$, the T observations can be compiled to give

$$
y = X\beta_0 + \varepsilon, \quad \text{where} \quad E(\varepsilon) = 0 \quad \text{and} \quad D(\varepsilon) = \Sigma. \tag{84}
$$

The remaining task of this section is to show that the information of $\{y_1, \ldots, y_t\}$ is also conveyed by the prediction errors or innovations $\{e_1, \ldots, e_t\}$

and that the latter are mutually uncorrelated random variables. For this purpose, consider substituting (68) and (70) into (73) to give

$$\begin{aligned} b_t &= \Phi_t b_{t-1} + K_t(y_t - H_t\Phi_t b_{t-1}) \\ &= \Lambda_t b_{t-1} + K_t y_t, \end{aligned} \tag{85}$$

where $\Lambda_t = (I - K_t H_t)\Phi_t$ is from (75). Repeated back-substitution gives

$$b_t = \Lambda_{t,1} b_0 + \sum_{j=1}^{t} \Lambda_{t,j+1} K_j y_j, \tag{86}$$

where $\Lambda_{t,j} = \Lambda_t \cdots \Lambda_j$ is a product of matrices that specialises to $\Lambda_{t,t} = \Lambda_t$ and to $\Lambda_{t,t+1} = I$. It follows that

$$\begin{aligned} e_t &= y_t - H_t\Phi_t b_{t-1} \\ &= y_t - H_t\Phi_t\Lambda_{t-1,1} b_0 - H_t\Phi_t \sum_{j=1}^{t-1} \Lambda_{t-1,j+1} K_j y_j, \end{aligned} \tag{87}$$

which is a straightforward generalisation of (53). On defining the vector $e = [e_1', \ldots, e_T']'$, the T equations can be written as

$$e = Ly - Wb_0 = L(y - Xb_0), \quad \text{with} \quad E(e) = 0 \quad \text{and} \quad D(e) = F. \tag{88}$$

Here, the matrix L is lower-triangular with units on the diagonal. The second equality follows from the fact that $E(e) = 0$ and $E(y) = Xb_0$, whence $Wb_0 = LXb_0$ for all b_0 and, therefore, $W = LX$.

Equation (87) shows that each error e_t is a linear function of $y_1, \ldots, y_t$. Next, we demonstrate that each y_t is a linear function of $e_1, \ldots, e_t$. By back-substitution in the equation $b_{t-1} = \Phi_{t-1} b_{t-2} + K_{t-1} e_{t-1}$, derived from (68) and (73), we get

$$b_{t-1} = \Phi_{t-1,1} b_0 + \sum_{j=1}^{t-1} \Phi_{t-1,j+1} K_j e_j. \tag{89}$$

Substituting $b_{t|t-1} = \Phi_t b_{t-1}$ into equation (70) gives

$$\begin{aligned} y_t &= H_t b_{t|t-1} + e_t \\ &= H_t\Phi_{t,1} b_0 + H_t \sum_{j=1}^{t-1} \Phi_{t,j+1} K_j e_j + e_t. \end{aligned} \tag{90}$$

Given that there is a one-to-one linear relationship between the observations and the prediction errors, it follows that we can represent the information set in terms of either. Thus, we have $\mathcal{I}_{t-1} = \{e_{t-1}, \ldots, e_1, \mathcal{I}_0\}$; and, given that $e_t = y_t - E(y_t|\mathcal{I}_{t-1})$, it follows from (A.8.vi) that e_t is uncorrelated with the preceding errors $e_1, \ldots, e_{t-1}$. The result indicates that the prediction errors are mutually uncorrelated.

8 Likelihood Functions and the Initial State Vector

Considerable attention has been focused by econometricians on the problem of estimating the initial state vector β_0 when the information concerning its distribution is lacking. This is a complicated matter that must be approached with care. The present section lays the necessary groundwork.

It has been assumed that the initial state vector has a normal prior distribution with $E(\beta_0) = b_0$ and $D(\beta_0) = P_0$. The sample data are generated by the equation $y = X\beta_0 + \varepsilon$ of (84), where the disturbances are normally distributed with $E(\varepsilon) = 0$ and $D(\varepsilon) = \Sigma$. Thus $E(y) = XE(\beta_0) + E(\varepsilon)$ and $D(y) = XD(\beta_0)X' + D(\varepsilon)$, so

$$E(y) = Xb_0, \tag{91}$$

$$D(y) = XP_0X' + \Sigma, \tag{92}$$

$$E(\beta_0) = b_0, \tag{93}$$

$$D(\beta_0) = P_0, \tag{94}$$

$$C(y, \beta_0) = XP_0. \tag{95}$$

The joint density function of y and β_0 is

$$N(y, \beta_0) = (2\pi)^{-(T+k)/2}|D(y, \beta_0)|^{-1/2} \exp\{-S(y, \beta_0)/2\}, \tag{96}$$

whose exponent, according to (A.6), can be written variously as

$$\begin{aligned} S(y, \beta_0) &= \begin{bmatrix} y - Xb_0 \\ \beta_0 - b_0 \end{bmatrix}' \begin{bmatrix} XP_0X' + \Sigma & XP_0 \\ P_0X' & P_0 \end{bmatrix}^{-1} \begin{bmatrix} y - Xb_0 \\ \beta_0 - b_0 \end{bmatrix} \\ &= \begin{bmatrix} y - E(y|\beta_0) \\ \beta_0 - b_0 \end{bmatrix}' \begin{bmatrix} \Sigma & 0 \\ 0 & P_0 \end{bmatrix}^{-1} \begin{bmatrix} y - E(y|\beta_0) \\ \beta_0 - b_0 \end{bmatrix} \\ &= \begin{bmatrix} y - Xb_0 \\ \beta_0 - E(\beta_0|y) \end{bmatrix}' \begin{bmatrix} XP_0X' + \Sigma & 0 \\ 0 & (X'\Sigma^{-1}X + P_0^{-1})^{-1} \end{bmatrix}^{-1} \begin{bmatrix} y - Xb_0 \\ \beta_0 - E(\beta_0|y) \end{bmatrix}. \end{aligned} \tag{97}$$

In the final expression, the identity

$$P_0 - P_0X'(XP_0X' + \Sigma)^{-1}XP_0 = (X'\Sigma^{-1}X + P_0^{-1})^{-1}, \tag{98}$$

which follows from (A.3.iii), has been used to obtain the expression for $D(\beta_0|y) = (X'\Sigma^{-1}X + P_0^{-1})^{-1}$.

In equation (97), there are two conditional expectations. The first, which is the mean of the conditional density function $N(y|\beta_0)$, is the familiar $E(y|\beta_0) = X\beta_0$. The second, which is the mean of $N(\beta_0|y)$, can be found by applying the regression formula (A.8.i). It is given by

$$
\begin{aligned}
E(\beta_0|y) &= b_0 + P_0X'(XP_0X' + \Sigma)^{-1}(y - Xb_0) \\
&= b_0 + (X'\Sigma^{-1}X + P_0^{-1})^{-1}X'\Sigma^{-1}(y - Xb_0) \\
&= (X'\Sigma^{-1}X + P_0^{-1})^{-1}(X'\Sigma^{-1}y + P_0^{-1}b_0) = b_*,
\end{aligned} \tag{99}
$$

where, to obtain the second expression, we have used the identity

$$
P_0X'(XP_0X' + \Sigma)^{-1} = (X'\Sigma^{-1}X + P_0^{-1})^{-1}X'\Sigma^{-1}. \tag{100}
$$

(This identity, which is in the form of $BD^{-1} = A^{-1}C$, can be converted to the form of $AB = CD$, from which it can be verified easily.)

Equation (97) can be written in a summary notation as

$$
\begin{aligned}
S(y, \beta_0) &= S(y|\beta_0) + S(\beta_0) \\
&= S(\beta_0|y) + S(y),
\end{aligned} \tag{101}
$$

where the following quadratic forms are from the exponents of the density functions $N(y|\beta_0)$, $N(\beta_0)$, $N(\beta_0|y)$ and $N(y)$ respectively:

$$
S(y|\beta_0) = (y - X\beta_0)'\Sigma^{-1}(y - X\beta_0), \tag{102}
$$

$$
S(\beta_0) = (\beta_0 - b_0)'P_0^{-1}(\beta_0 - b_0), \tag{103}
$$

$$
S(\beta_0|y) = (\beta_0 - b_*)'(X'\Sigma^{-1}X + P_0^{-1})(\beta_0 - b_*), \tag{104}
$$

$$
\begin{aligned}
S(y) &= (y - Xb_0)'(XP_0X' + \Sigma)^{-1}(y - Xb_0) \\
&= (y - Xb_0)'\{\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X + P_0^{-1})^{-1}X'\Sigma^{-1}\}(y - Xb_0).
\end{aligned} \tag{105}
$$

The second expression for $S(y)$ on the RHS of (105) follows from (A.3.iii). There is also a relationship $|D(y, \beta_0)| = |D(y|\beta_0)||D(\beta_0)| = |D(\beta_0|y)||D(y)|$ relating the determinantal terms of the various distributions, which gives rise to the identity

$$
|P_0| = |XP_0X' + \Sigma||X'\Sigma^{-1}X + P_0^{-1}|^{-1}. \tag{106}
$$

The various ways for estimating β_0 can be considered in the light of the foregoing algebraic results. First, the estimator can be obtained by maximising, in respect of β_0, the likelihood function corresponding to the conditional density function $N(y|\beta_0)$. In this approach, β_0 tends to be regarded as a parametric constant, rather the realised value of a random variable, so that the *conditional* likelihood function becomes *unconditional*. In any event, the result obtained by minimising the quadratic function $S(y|\beta_0)$ of (108), will be described as the unconditional full-sample estimator:

$$
b_{0|T} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y. \tag{107}
$$

Substituting this into $N(y|\beta_0)$ gives the concentrated function

$$
N^c(y) = (2\pi)^{-T/2}|\Sigma|^{-1/2}\exp\{-S^c(y)/2\}, \tag{108}
$$

wherein

$$S^c(y) = y'\{\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\}y. \tag{109}$$

The concentrated function provides a criterion function from which to derive the maximum-likelihood estimates of the fundamental system parameters that are to be found within H_t, Φ_t, Ω_t and H_t.

Next, consider the estimator of the initial state vector determined by the conditional expectation $b_* = E(\beta_0|y)$, specified in alternative forms by (99). This estimator can also be derived by minimising $S(y, \beta_0) = S(y|\beta_0) + S(\beta_0)$ in respect of β_0 according to the principle of mixed estimation, which is equivalent to maximising the likelihood function corresponding to the joint density function $N(y, \beta_0)$. By, letting $P_0 \to \infty$ in (99), which is tantamount to negating the priori information on β_0, we get the unconditional estimator $b_{0|T}$ of (107), as one might expect.

In the absence of informative prior information, we can also attempt to obtain an estimate of $E(\beta_0) = b_0$ from the likelihood function corresponding to the marginal density function

$$N(y) = (2\pi\sigma)^{-T/2}|XP_0X' + \Sigma|^{-1/2}\exp\{-S(y)/2\}, \tag{110}$$

wherein the quadratic exponent $S(y)$ is given by (105). Differentiating $S(y)$ with respect to b_0 and setting the result to zero gives a first-order condition from which to obtain the maximum-likelihood estimator

$$\begin{aligned}\hat{b}_0 &= \{X'(XP_0X' + \Sigma)^{-1}X\}^{-1}X'(XP_0X' + \Sigma)^{-1}y \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y = b_{0|T}.\end{aligned} \tag{111}$$

The second expression, which is just the unconditional estimator of β_0, follows from the result on equivalent regression metrics. This result indicates that the generalised least-squares estimators of β in the regression models $(y; X\beta, \Omega_1)$ and $(y; X\beta, \Omega_2)$ are identical if and only if the columns of the matrices $\Omega_1^{-1}X$ and $\Omega_2^{-1}X$ span the same space—see Pollock (1979, p. 86). The equality can be demonstrated directly by reference to (100), which gives $X'(XP_0X' + \Sigma)^{-1} = P_0^{-1}(X'\Sigma^{-1}X + P_0^{-1})X'\Sigma^{-1}$. After substituting this in the first expression on the RHS of (111), the factors P_0^{-1} and $(X'\Sigma^{-1}X + P_0^{-1})$ can be cancelled with their inverses to give the second expression.

Setting $b_0 = b_{0|T}$ in the marginal density function gives a concentrated likelihood function whose quadratic exponent is $S^c(y)$ of (109). This can be seen via the second expression of (105). The likelihood can be maximised further by setting $P_0 = 0$. The result is, once more, the function $N^c(y)$ of (108). Setting $P_0 = 0$ is an unnatural recourse in circumstances where there is no prior information regarding β_0. However, it accords with the fact that the dispersion of the estimate $b_{0|T}$ is a function of sample information alone.

Finally, we should allow $P_0 \to \infty$ within the marginal distribution $N(y)$ of (110) which will set $S(y) \to S^c(y)$ in the exponent. This creates what is

described in de Jong (1988a, 1991) and Ansley and Kohn (1985a, 1990) as a diffuse distribution. Taking limits within the determinantal term is problematic, since XP_0X' is unbounded. However, in view of (106), the term can be written as $|XP_0X' + \Sigma|^{-1/2} = |P_0|^{-1/2}|X'\Sigma^{-1}X + P_0^{-1}|^{-1/2}$. Therefore, it has been proposed by de Jong to omit the factor $|P_0|^{-1/2}$ and define the diffuse likelihood function by

$$N^d(y) = |X'\Sigma^{-1}X|^{-1/2}(2\pi)^{-T/2}\exp\{-S^c(y)/2\}. \tag{112}$$

The exponent $S^c(y)$ of the diffuse likelihood, which is the essential part, is identical to that arising from concentrating the marginal likelihood function $N(y)$ of (110) in respect of b_0 and P_0 or, equally, from concentrating the conditional likelihood function $N(y|\beta_0)$ in respect of β_0.

It is arguable that negating the prior information by letting $P_0 \to \infty$ is best done in the context of the joint distribution factorised as $N(y, \beta_0) = N(y|\beta_0)N(\beta_0)$. This confines the difficulties of the limiting process to the factor $N(\beta_0)$.

Example. There are several alternative ways for deriving the quadratic component of the marginal distribution $N(y)$ that lead to expressions so different that it is difficult demonstrate their equivalence.

Setting $\beta_0 = E(\beta_0|y) = b_*$ within the exponent $S(y, \beta_0) = S(\beta_0|y) + S(y)$ of the product $N(y, \beta_0) = N(\beta_0|y)N(y)$ gives $S(y)$, since the term $S(\beta_0|y)$ is thereby eliminated. This result holds true however the expression for $S(y, \beta_0)$ is derived. Thus, setting $\beta_0 = b_*$ in $S(y, \beta_0) = S(\beta_0) + S(y|\beta_0)$ gives

$$S(y) = (b_* - b_0)'P_0^{-1}(b_* - b_0) + (y - Xb_*)'\Sigma^{-1}(y - Xb_*). \tag{113}$$

This expression has been exploited by Goméz and Maravall (1994), and the same procedure has been followed in Box and Jenkins (1976) in finding the "unconditional sum of squares" of an ARMA model.

An alternative route to the marginal distribution is via the identity $N(y) = N(y|\beta_0)N(\beta_0)/N(\beta_0|y)$. This leads to $S(y) = S(y|\beta_0) + S(\beta_0) - S(\beta_0|y)$, which becomes

$$\begin{aligned} S(y) = &(y - X\beta_0)'\Sigma^{-1}(y - X\beta_0) + (\beta_0 - b_0)'P_0^{-1}(\beta_0 - b_0) \\ &-(\beta_0 - b_*)'(X'\Sigma^{-1}X + P_0^{-1})(\beta_0 - b_*). \end{aligned} \tag{114}$$

After expanding the quadratics, the terms in β_0 can be cancelled from this expression. This formulation has been employed by de Jong (1988a, 1991).

When either (113) or (114) are used as the criterion function for estimating b_0, the functional dependence of $b_* = E(\beta_0|y)$ on b_0 must be taken into account.

9 Calculating the Estimate of the Initial State

There are various practical means for obtaining the values of $\mathcal{I}_0 = \{b_0, P_0\}$ to start the Kalman filter. Often analytic expressions for b_0 and P_0 can be found by assuming that the state vectors are generated by a stationary process. Then, the matrices H_t, Φ_t, Ω_t and Ψ_t become constant and loose their temporal subscripts.

For stationarity, the eigenvalues of the transformation matrix Φ must lie within the unit circle, which implies that $\lim(n \to \infty)\Phi^n = 0$. Then, the unconditional moments $E(\beta_0) = b_0 = 0$ and $D(\beta_0) = P_0 = \Phi P_0 \Phi' + \Psi$ from (66) provide the starting values. The initial dispersion matrix can be found by calculating $P_0 = (I - \Phi \otimes \Phi)^{-1}\text{vec}\Psi$ via a matrix inversion. Alternatively, it can be found via a convergent iterative process whose ith step is described by $P_i = \Phi P_{i-1}\Phi' + \Psi$.

When the state space equations (65) and (66) represent an ARMA process, there are well-known methods for finding the autocovariances of the process that can be used in forming P_0—see Pollock (1999). The state-space representation of the ARMA model may be formulated to facilitate the direct derivation of P_0, as in Mittnik (1987a, 1987b) and Diebold (1986a, 1986b).

In the econometric literature, there is a tendency to adopt the transformations approach to initialise the Kalman filter when it is applied to a nonstationary process. This reflects the influence of Ansley and Kohn (1985a). The purpose of the transformation is to eliminate the dependence of the likelihood upon unknown initial values with a diffuse or improper distribution.

The transformations approach can cause confusion when it is used as a theoretical device with no intended application. Indeed, the modified Kalman filter of Ansley and Kohn (1985a) is designed to avoid transformations of the data that obstruct the handling of the problem of missing observations.

To illustrate the theoretical approach of Ansley and Kohn, consider the orthonormal matrix $C = [C_1, C_2]$ defined in Section 5 in connection with the LUS residuals. The columns of C_1 span the same space as those of X, whereas $C_2'X = 0$. Therefore, transforming the equation $y = X\beta_0 + \varepsilon$ of (84) by C' gives

$$\begin{bmatrix} C_1'y \\ C_2'y \end{bmatrix} = \begin{bmatrix} C_1'X\beta_0 \\ 0 \end{bmatrix} + \begin{bmatrix} C_1'\varepsilon \\ C_2'\varepsilon \end{bmatrix}, \tag{115}$$

where $D(C_2'y) = C_2'\Sigma C_2$. The likelihood function of $C_2'y$ embodies the concentrated sum of squares

$$S^c(y) = y'C_2(C_2'\Sigma C_2)^{-1}C_2'y = y'\{\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\}y. \tag{116}$$

The second equality of (116) follows from the fact that, if $\text{Rank}[W, X] = T$

and if $W'\Sigma^{-1}X = 0$, then

$$W(W'\Sigma^{-1}W)^{-1}W'\Sigma^{-1} = I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}. \tag{117}$$

The equality is obtained by premultiplying both sides of (117) by Σ^{-1} and then setting $W = \Sigma C_2$.

The expression for $S^c(y)$ on the RHS of (116), which is also given by (109), can be obtained by replacing β_0 in $S(y|\beta_0) = (y - X\beta_0)'\Sigma^{-1}(y - X\beta_0)$ by the full-sample estimate $b_{0|T}$ of (107). Equally, it can be obtained from $S(y)$ of (105) by setting $b_0 = b_{0|T}$. Observe that, when $\Sigma = I$, equation (116) specialises to (58), which represents the sum of squares of the LUS residuals of the ordinary regression model.

To fulfil the conditions of (115) and (116), C does not have to be an orthonormal matrix. An alternative transformation, which can be used in practice, has been proposed in Bell and Hillmer (1991) in the context of their treatment of the unobserved components model. They set $X = [X_1', X_2']'$ and $y = [y_1', y_2']'$, where X_1 and y_1 comprise the first k observations and k is the dimension of β_0. Then, they form

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} X_1^{-1} & 0 \\ -X_2X_1^{-1} & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ 0 \end{bmatrix} + \begin{bmatrix} X_1^{-1}\varepsilon_1 \\ -X_2X_1^{-1}\varepsilon_1 + \varepsilon_2 \end{bmatrix}. \tag{118}$$

Here, $X_1^{-1}y_1 = b_{0|k}$ is an estimator of β_0 based on minimal data, whilst

$$S^c(y) = z_2'D^{-1}(z_2)z_2 = (y_2 - X_2b_{0|k})'D^{-1}(z_2)(y_2 - X_2b_{0|k}) \tag{119}$$

is an alternative representation of the concentrated sum of squares. This expression is analogous to (56), which relates to ordinary recursive regression in the absence of prior information. One should note that, if $D(\varepsilon) = \Sigma = I$, then $D(z_2) = X_2(X_1'X_1)^{-1}X_2' + I$, which would make the RHS of (119) identical to (56). Observe that, if $C_1' = [X_1^{-1}, 0]$ and $C_2' = [-X_2X_1^{-1}, I]$, then $C_1'y = z_1$ and $C_2'y = z_2$ satisfy (115) and (116).

Equations (116) and (119) represent the same quantity; and comparing them shows that the concentrated function may be expressed in terms of the minimal estimate $b_{0|k}$ as well as the full-sample estimate $b_{0|T}$. This seeming paradox, which is analogous to a feature of ordinary recursive regression, described in Section 5, points to two ways of handling the start-up problem.

We shall begin by describing a procedure that incorporates the full-sample estimates of the start-up values. Then we shall show how the procedure can be modified to incorporate the minimal estimates, which are repeatedly enhanced as the data are assimilated during the process of the recursive estimation.

Consider the following expression for the quadratic function within $N(y)$:

$$\begin{aligned} S(y) &= (y - Xb_0)'(XP_0X' + \Sigma)^{-1}(y - Xb_0) \\ &= (y - Xb_0)'L'F^{-1}L(y - Xb_0) = e'F^{-1}e, \end{aligned} \tag{120}$$

where F is a block-diagonal matrix with F_t as the tth diagonal block. Here, the first expression on the RHS is from (105), whereas the second expression, which reflects the identities of (88), is the form proposed originally by Schweppe (1965).

It has been show, in Section 8, that the value that minimises $S(y)$ is the estimator $b_{0|T} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ of (107) and (111), which is invariant with respect to the value of P_0. Therefore, in estimating b_0, one is liable to set $P_0 = 0$, which is tantamount to replacing the marginal function $S(y)$ by the conditional function $S(y|\beta_0) = (y - X\beta_0)'\Sigma^{-1}(y - X\beta_0)$ of (102). (Setting $P_0 = 0$, in this context does not carry the literal interpretation that β_0 is now known with certainty. Nor should it convey the usual interpretation that β_0 is to be regarded as a "constant". The only reasonable interpretation is that it signals a replacement of the marginal function by the conditional function.)

The form of the estimator $b_{0|T}$ given under (107) is not computable. To derive an operational form, consider writing equation (87) as

$$\begin{aligned} e_t &= \Big\{y_t - H_t\Phi_t \sum_{j=1}^{t-1} \Lambda_{t-1,j+1}K_j y_j\Big\} - H_t\Phi_t\Lambda_{t-1,1}b_0 \\ &= e_t^* - W_t b_0, \end{aligned} \tag{121}$$

where e_t^* and $W_t b_0$ are the tth subvectors, respectively, of Ly and $Wb_0 = LXb_0$, which are to be found in equation (88). Substituting in $S(y) = \sum_{t=1}^{T} e_t' F^{-1} e_t$, which is the final expression from (120), gives

$$S(y) = \sum_{t=1}^{T} (e_t^* - W_t b_0)' F_t^{-1} (e_t^* - W_t b_0). \tag{122}$$

The estimated starting value, obtained by minimising this in respect of b_0, is

$$b_{0|T} = \Big(\sum_{t=1}^{T} W_t' F_t^{-1} W_t\Big)^{-1} \sum_{t=1}^{T} W_t' F_t^{-1} e_t^* = M_T^{-1} m_T. \tag{123}$$

The elements of this expression can be accumulated via the recursions

$$\begin{aligned} m_t &= m_{t-1} + \Lambda_{t-1,1}'\Phi_t' H_t' F_t^{-1} e_t^*, \\ M_t &= M_{t-1} + \Lambda_{t-1,1}'\Phi_t' H_t' F_t^{-1} H_t \Phi_t \Lambda_{t-1,1}, \end{aligned} \tag{124}$$

which begin with $m_0 = 0$, $M_0 = 0$. They should be run parallel to the Kalman filter initialised with $b_0 = 0$ and $P_0 = 0$. To accumulate $\Lambda_{t-1,1}$, we can define a recursion

$$\Lambda_{t,1} = (\Phi_t - K_t H_t \Phi_t)\Lambda_{t-1,1}, \tag{125}$$

which starts with $\Lambda_{1,1} = \Lambda_1$. Notice, however, in reference to (121), that the requisite quantities can be obtained by exploiting the recursion that gives rise

to the sequence of prediction errors. By starting that recursion with $b_0 = 0$, the sequence $\{e_t^*\}$ is generated instead of the sequence $\{e_t = e_t(b_0)\}$. By replacing b_0 by an identity matrix and the observations y_t by zeros, the sequence $\{W_t\}$ is generated.

The estimation of the initial conditions can, therefore, be accomplished by extending two of the equations of the Kalman filter and by adding an extra one:

$$E_t = Y_t - H_t\Phi_t B_{t-1}, \qquad \textit{Extended Prediction Error} \tag{126}$$

$$B_t = \Phi_t B_{t-1} + K_t E_t, \qquad \textit{Extended State Estimate} \tag{127}$$

$$G_t = G_{t-1} + E_t' F_t^{-1} E_t. \qquad \textit{Cross - Product Accumulation} \tag{128}$$

Here, (126) and (127) are extensions of (70) and (73), respectively. The matrices $E_t = [e_t^*,\ W_t]$ and $B_t = [b_t^*,\ \Lambda_{t,1}]$ have the prediction error and the state estimate of the ordinary Kalman filter (assuming a starting value of $b_0 = 0$) in their leading columns, respectively, whilst $Y_t = [y_t,\ 0]$. The starting values of the extended filter are $B_0 = [0,\ I]$, $P_0 = 0$ and $G_0 = 0$. The matrix G_t is as follows:

$$G_t = \begin{bmatrix} S_t & m_t \\ m_t' & M_t \end{bmatrix}. \tag{129}$$

This contains the quantities defined in (124) together with the sum of squares of the prediction errors scaled by their variance.

The algorithm is attributable to Rosenberg (1973). It is expounded in Harvey (1989), and elsewhere, and it is used in de Jong (1988a, 1988b, 1989, 1991a, 1991b). The procedure of Rosenberg was to generate the full sequence of state estimates $b_1^*, \ldots, b_T^*$ on the basis of the starting value $b_0 = 0$ and then to adjust them using the estimate $b_{0|T}$ of (123). It follows from (86) that the adjusted estimate of β_t is $b_t = b_t^* + \Lambda_{t,1} b_{0|T}$.

An alternative procedure, described by de Jong (1991a, 1991b), which is in accordance with the prescriptions of Bell and Hillmer (1991), is to collapse the extended filter at the earliest opportunity by absorbing the minimal estimate $b_{0|k} = M_k^{-1} m_k$ of the starting value into the state estimate. Then, $e_k = e_k^* - W_k M_k^{-1} m_k$ and $b_k = b_k^* + \Lambda_{k,1} M_k^{-1} m_k$ can be formed. The succeeding prediction errors and state estimates have values given by

$$\begin{aligned} e_t &= e_t^* - W_t M_t^{-1} m_t, \\ b_t &= b_t^* + \Lambda_{t,1} M_t^{-1} m_t, \end{aligned} \tag{130}$$

if one were to calculate the quantities on the RHS. Thus, the standard Kalman filter implicitly enhances the estimate of the initial state as the iterations proceed, but the enhanced estimate itself will not be available. The dispersion of the state estimate is

$$
\begin{aligned}
D(b_t) &= D(b_t^*) + \Lambda_{t,1} D(b_{0|t}) \Lambda_{t,1}' \\
&= P_t^* + \Lambda_{t,1} M_t^{-1} \Lambda_{t,1}' = P_t,
\end{aligned}
\tag{131}
$$

which is generated directly by the standard (collapsed) filter—see de Jong and Chu-Chun-Lin (1994).

A problem that may arise from collapsing filter is how to estimate the state vectors $\beta_1, \ldots, \beta_{k-1}$, that predate the collapse at $t = k$, when the first estimate of the starting value is formed. One solution, outlined in de Jong and Chu-Chun-Lin (2003), uses the estimate $b_{0|k} = M_k^{-1} m_k$ to adjust the pre-collapse values just as $b_{0|T} = M_T^{-1} m_T$ is used in Rosenberg's procedure. The resulting state estimates will be enhanced in a subsequent smoothing operation. In the case of the unobserved components model, the start-up values are the initial values of the component sequences; and they coincide with the elements of the initial state vector. Therefore, the problem does not arise.

The smoothed estimates of the state vectors are unaffected by whether $b_{0|k}$ or $b_{0|T}$ has been used in preliminary estimates obtained from filtering. Smoothing adds information that is missing from the estimates, but it has no effect if the information has already been incorporated.

The essence of a different method for initialising the filter due to Ansley and Kohn (1985a) has been presented already in Section 4 in the context of an ordinary recursive regression. This requires setting $P_t = P_t^* + \rho P_t^\circ$, where P_t° relates to the diffuse component of the prior information and where $\rho \to \infty$. When $P_t^\circ > 0$ and $f_t^\circ > 0$, the algorithm is summarised by equations (28), (30) and (31). When $P_t^\circ = 0$ and, therefore, $f_t^\circ = 0$, these are replaced by the corresponding equations of the standard algorithm.

Some minor elaborations are required to apply the method in the present context. First, we have $P_{t|t-1} = P_{t|t-1}^* + \rho P_{t|t-1}^\circ$, where

$$
P_{t|t-1}^\circ = \Phi_t P_{t-1}^\circ \Phi_t' \qquad \text{and} \qquad P_{t|t-1}^* = \Phi_t P_{t-1}^* \Phi_t' + \Psi_t. \tag{132}
$$

Then, the components of the prediction-error dispersion $F_t = F_t^* + \rho F_t^\circ$ must be defined:

$$
F_t^\circ = H_t P_{t|t-1}^\circ H_t' \qquad \text{and} \qquad F_t^* = H_t P_{t|t-1}^* H_t' + \Omega_t. \tag{133}
$$

Usually, one can assume that, when it is nonzero, F_t° is nonsingular—see Durbin and Koopman (2001). In the process of initialisation, when $P_t^\circ > 0$ and $F_t^\circ > 0$, the following equations are employed:

$$
b_t = b_{t|t-1} + P_{t|t-1}^\circ H_t' F_t^{\circ -1} (y_t - H_t b_{t|t-1}), \tag{134}
$$

$$
P_t^\circ = P_{t|t-1}^\circ - P_{t|t-1}^\circ H_t' F_t^{\circ -1} H_t P_{t|t-1}^\circ, \tag{135}
$$

$$
\begin{aligned}
P_t^* = {} & P_{t|t-1}^* + P_{t|t-1}^\circ H_t' F_t^{\circ -1} F_t^* F_t^{\circ -1} H_t P_{t|t-1}^\circ \\
& - P_{t|t-1}^\circ H_t' F_t^{\circ -1} H_t P_{t|t-1}^* - P_{t|t-1}^* H_t' F_t^{\circ -1} H_t P_{t|t-1}^\circ.
\end{aligned}
\tag{136}
$$

When the initialisation is complete, the conditions $F_t^\circ = 0$ and $P_t^\circ = 0$ prevail, and the equations above are replaced by

$$b_t = b_{t|t-1} + P^*_{t|t-1}H_t'F_t^{*-1}(y_t - H_t b_{t|t-1}), \tag{137}$$

$$P_t^\circ = P^\circ_{t|t-1}, \tag{138}$$

$$P_t^* = P^*_{t|t-1} - P^*_{t|t-1}H_t'F_t^{*-1}H_t P^*_{t|t-1}. \tag{139}$$

These are just the equations of the standard Kalman filter.

On defining

$$K_t^\circ = P^\circ_{t|t-1}H_t'F_t^{\circ-1} \qquad \text{and} \qquad \Lambda_t^\circ = (I - K_t^\circ H_t)\Phi_t, \tag{140}$$

we can write the equations (134), (135) and (136) as

$$b_t = \Lambda_t^\circ b_{t-1} + K_t^\circ y_t, \tag{141}$$

$$P_t^\circ = (I - K_t^\circ H_t)P^\circ_{t|t-1} = (I - K_t^\circ H_t)P^\circ_{t|t-1}(I - H_t'K_t'^\circ), \tag{142}$$

$$P_t^* = (I - K_t^\circ H_t)P^*_{t|t-1}(I - H_t'K_t'^\circ) + K_t^\circ \Omega_t K_t'^\circ. \tag{143}$$

The original derivation in Ansley and Kohn (1985a) is somewhat laborious, and a subsequent abbreviated derivation in Kohn and Ansley (1986) is more accessible. The use of the algorithm in estimating nonstationary ARMA models has been described in Ansley and Kohn (1985b) and Kohn and Ansley (1986). A modified version of the algorithm, claiming superior numerical accuracy, is provided in Ansley and Kohn (1990). Other derivations are given in Snyder (1988), which considers a square-root version of the Kalman filter, and in Koopman (1997) which treats the most general case, where $F_t^\circ > 0$ is not necessarily nonsingular.

One virtue of the foregoing method for initialising the filter is that it provides a complete sequence of state estimates and their corresponding dispersion matrices for $t = 1, \ldots, T$ that is amenable to a modified or supplemented version of the smoothing algorithm.

10 The Smoothing Algorithms

The Kalman filter, used as a real-time or on-line algorithm, estimates of the state vectors from current and past information. Often, it is possible to enhance these estimates using subsequent information.

In processing speech digitally, before its transmission by telephone, it is acceptable to impose a small delay for gathering extra information. A fixed-lag smoothing algorithm can then be used to enhance the digital signal. In econometrics, with no immediate real-time constraint, all the subsequent information within a given sample can be used to enhance the state estimates via the so-called fixed-interval smoothing algorithms.

Smoothing algorithms quickly followed the publication of Kalman (1960). A notable contribution is Rauch (1963), and the early work is surveyed in Meditch (1973). Whereas the fixed-lag smoothing algorithms feature prominently in the engineering literature, fixed-interval algorithms have received less attention; and econometricians have found scope for developing them. Notable contributions are Ansley and Kohn (1982), Kohn and Ansley (1989), de Jong (1988b, 1989) and Koopman (1993). All classes of smoothing algorithms are surveyed and compared in Merkus, Pollock and de Vos (1993).

This section concentrates exclusively on the fixed-interval algorithms, taking the sequence of prediction errors $\mathcal{I}_T = \{e_1, \ldots, e_T\}$ to represent the information set. Because the prediction errors are mutually independent, (A.8.i) implies that

$$E(\beta_t|\mathcal{I}_T) = E(\beta_t|\mathcal{I}_t) + \sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)e_j. \tag{144}$$

This indicates how the estimate $b_t = E(\beta_t|\mathcal{I}_t)$ is updated using the information $\{e_{t+1}, \ldots, e_T\}$, which has arisen after time t, to produce the definitive estimate $b_{t|T} = E(\beta_t|\mathcal{I}_T)$. According to (A.8.ii), the dispersion matrix is

$$D(\beta_t|\mathcal{I}_T) = D(\beta_t|\mathcal{I}_t) - \sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)C(e_j, \beta_t). \tag{145}$$

To realise these equations, we need a computationally efficient recursion.

Consider

$$e_k = H_k\Phi_k(\beta_{k-1} - b_{k-1}) + H_k\nu_k + \eta_k, \tag{146}$$

which comes from substituting the transition equation (66) into the observation equation (65) to give $y_k = H_k(\Phi_k\beta_{k-1} + \nu_k) + \eta_k$ and then subtracting $H_k b_{k|k-1} = H_k\Phi_k b_{k-1}$. Within this expression, there is

$$\beta_{k-1} - b_{k-1} = \Lambda_{k-1}(\beta_{k-2} - b_{k-2}) + (I - K_{k-1}H_{k-1})\nu_{k-1} - K_{k-1}\eta_{k-1}. \tag{147}$$

This is obtained by subtracting $b_{k-1} = \Phi_{k-1}b_{k-2} + K_{k-1}e_{k-1}$ from the transition equation and then substituting the expression for e_{k-1} from (146) into the result. The equation is amenable to recursion, running from $k-1$ down to t, which gives

$$\beta_{k-1} - b_{k-1} = \Lambda_{k-1,t+1}(\beta_t - b_t) + \sum_{j=t+1}^{k-1} \Lambda_{k-1,j+1}\{(I - K_jH_j)\nu_j - K_j\eta_j\}. \tag{148}$$

The summation comprises stochastic elements that are subsequent to t and, therefore, independent of the prediction error e_t. After incorporating (148) in (146), it follows, when $k > t$, that

$$C(\beta_t, e_k) = E\{\beta_t(\beta_t - b_t)'\Lambda'_{k-1,t+1}\Phi'_k H'_k\} \tag{149}$$
$$= P_t\Lambda'_{k-1,t+1}\Phi'_k H'_k.$$

Now consider

$$C(\beta_{t+1}, e_k) = P_{t+1}\Lambda'_{k-1,t+2}\Phi'_k H'_k. \tag{150}$$

Comparing (149) and (150) shows that

$$C(\beta_t, e_k) = P_t\Lambda'_{t+1}P^{-1}_{t+1}C(\beta_{t+1}, e_k) \tag{151}$$
$$= P_t\Phi'_{t+1}P^{-1}_{t+1|t}C(\beta_{t+1}, e_k).$$

Here, the identity $P^{-1}_{t+1}\Lambda_{t+1} = P^{-1}_{t+1|t}\Phi_{t+1}$, giving the second equality, comes via (74) and (75), which indicate that $P_{t+1} = \Lambda_{t+1}\Phi^{-1}_{t+1}P_{t+1|t}$. Equation (151) provides the recursion to implement the formulae of (144) and (145). The classical fixed-interval smoother is derived from

$$E(\beta_t|\mathcal{I}_T) = E(\beta_t|\mathcal{I}_t) + P_t\Phi'_{t+1}P^{-1}_{t+1|t}\sum_{j=t+1}^{T} C(\beta_{t+1}, e_j)D^{-1}(e_j)e_j, \tag{152}$$

which is obtained by substituting the identity of (151) into (144). But

$$E(\beta_{t+1}|\mathcal{I}_T) = E(\beta_{t+1}|\mathcal{I}_t) + \sum_{j=t+1}^{T} C(\beta_{t+1}, e_j)D^{-1}(e_j)e_j, \tag{153}$$

so it follows that (152) can be rewritten as

$$b_{t|T} = b_t + P_t\Phi'_{t+1}P^{-1}_{t+1|t}\{b_{t+1|T} - b_{t+1|t}\}, \tag{154}$$

where $b_{t+1|T} = E(\beta_{t+1}|\mathcal{I}_T)$ and $b_{t+1|t} = E(\beta_{t+1}|\mathcal{I}_t)$ have been used for conciseness. This is the classical formula for the fixed-interval smoother.

A similar strategy can be used to derive the dispersion matrix of the smoothed estimate. Corresponding to (153), we have

$$D(\beta_{t+1}|\mathcal{I}_T) = D(\beta_{t+1}|\mathcal{I}_t) - \sum_{j=t+1}^{T} C(\beta_{t+1}, e_j)D^{-1}(e_j)C(e_j, \beta_{t+1})e_j. \tag{155}$$

Therefore, (145) can be written as

$$P_{t|T} = P_t - P_t\Phi'_{t+1}P^{-1}_{t+1|t}\{P_{t+1|T} - P_{t+1|t}\}P^{-1}_{t+1|t}\Phi_{t+1}P_t. \tag{156}$$

The classical formulae presuppose a sequence $b_t; t = 1, \ldots, T$ of state estimates generated by forward filtering. Smoothing is effected by running backward through the sequence using a first-order feedback in respect of the smoothed estimates. The algorithm is due to Rauch (1963) and its derivation can be found in Anderson and Moore (1979) amongst other sources.

In circumstances where $P_t\Phi'_{t+1}P^{-1}_{t+1|t}$ can be represented by a constant matrix, the classical algorithm is efficient and easy to implement. This occurs if there is constant transition matrix Φ and if the filter gain K_t converges to a constant. In all other circumstances, it is necessary recompute the factor at each iteration and the algorithm is liable to cost time and invite numerical inaccuracies. The burden of inverting of $P_{t+1|t}$ can be avoided at the expense of generating a supplementary sequence to accompany the smoothing process.

Consider the summation within (144), which, using (149), can be written as

$$\begin{aligned} &\sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)e_j \\ &= P_t \sum_{j=t+1}^{T} \Lambda'_{j-1,t+1}\Phi'_j H'_j F_j^{-1} e_j = P_t q_{t+1}. \end{aligned} \tag{157}$$

Within (145), there is also

$$\begin{aligned} &\sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)C(\beta_t, e_j) \\ &= P_t\Big\{ \sum_{j=t+1}^{T} \Lambda'_{j-1,t+1}\Phi'_j H'_j F_j^{-1} H_j \Phi_j \Lambda_{j-1,t+1}\Big\}P_t = P_t Q_{t+1} P_t. \end{aligned} \tag{158}$$

Here, the terms q_{t+1} and Q_{t+1} are elements of sequences generated by recursions running backwards in time that take the form

$$\begin{aligned} q_t &= \Phi'_t H'_t F_t^{-1} e_t + \Lambda'_{t+1} q_{t+1}, \\ Q_t &= \Phi'_t H'_t F_t^{-1} H'_t \Phi'_t + \Lambda'_{t+1} Q_{t+1} \Lambda_{t+1}, \end{aligned} \tag{159}$$

and that are initiated with $q_T = \Phi'_T H'_T F_T^{-1} e_T$ and $Q_T = \Phi'_T H'_T F_T^{-1} H_T \Phi_T$. These are the counterparts of the recursions of (124) that run forwards in time. The recursions of (159) provide an alternative to the classical fixed-interval smoothing algorithm. Thus, putting (157) and (158) into (144) and (145), respectively, gives

$$\begin{aligned} b_{t|T} &= b_t + P_t q_{t+1}, \\ P_{t|T} &= P_t - P_t Q_{t+1} P_t. \end{aligned} \tag{160}$$

This algorithm is due to de Jong (1989).

The smoothing algorithms can be adapted to take account of diffuse initial conditions. Let $t = k$ be the point where there is just sufficient sample information to determine unique state estimates. This is the point at which the diffuse filter makes its transition to the standard form. Then, for $t < k$, we have $P_t = P_t^* + \rho P_t^\circ$ and $F_t = F_t^* + \rho F_t^\circ$. The latter gives rise to

$$F_t^{-1} = (F_t^* + \rho F_t^\circ)^{-1} \tag{161}$$

$$
\begin{aligned}
&= \rho^{-1}(F_t^\circ)^{-1} - \rho^{-2}(F_t^\circ)^{-1}F_t^*(F_t^\circ)^{-1} + \cdots \\
&= \rho^{-1}(F_t^\circ)^{-1} + O(\rho^{-2}),
\end{aligned}
$$

and to

$$
\begin{aligned}
\Lambda_t &= (I - K_tH_t)\Phi_t = (I - K_t^\circ H_t)\Phi_t - \rho^{-1}K_t^* H_t\Phi_t + O(\rho^{-2}), \\
&= \Lambda_t^\circ + O(\rho^{-1}),
\end{aligned} \tag{162}
$$

where $K_t^\circ = P_{t|t-1}^\circ H_t F_t^{\circ -1}$ and $\Lambda_t^\circ = (I - K_tH_t)\Phi_t$ are defined in (140) and where

$$
K_t^* = P_{t-1}^* H_t (F_t^\circ)^{-1} - P_{t-1}^\circ H_t (F_t^\circ)^{-1} F_t^* (F_t^\circ)^{-1}. \tag{163}
$$

These results follow analogously to those Section 4.

Now consider the expression

$$
C(\beta_t, e_j)D^{-1}(e_j)e_j = (P_t^* + \rho P_t^\circ)\Lambda_{j-1,t+1}'\Phi_j' H_j' F_j^{-1} e_j, \tag{164}
$$

which is found in the formula (144) for the fixed-interval smoother in the case when $t < k$ if there are diffuse initial conditions. First, in view of (162) and (163), we find that, when $\rho \to \infty$,

$$
P_t^* \Lambda_{j-1,t+1}'\Phi_j' H_j' F_j^{-1} = 0, \qquad \text{when} \quad t < k. \tag{165}
$$

When $j \geq k$, $P_t^\circ \Lambda_{j-1,t+1}^{\circ\prime} = P_0^\circ \Lambda_{j-1,1}^{\circ\prime} = 0$. Thus, when $\rho \to \infty$,

$$
\rho P_t^\circ \Lambda_{j-1,t+1}'\Phi_j' H_j' F_j^{-1} = \begin{cases} 0, & \text{if } t < k \leq j, \\ P_t^\circ \Lambda_{j-1,t+1}^{\circ\prime}\Phi_j' H_j' (F_j^\circ)^{-1}, & \text{if } t < j < k. \end{cases} \tag{166}
$$

Recognising these conditions, we can extend the algorithm (159) and (160) to the case of diffuse initial conditions. The standard recursion, indicated by (159), runs from $t = T$ down to $t = k$ and generates values that may be denoted by q_t^*. Thereafter, from $t = k - 1$ down to $t = 1$, the values of this sequence, together with the values q_t° of a supplementary sequence, beginning with $q_k^\circ = 0$, are generated by the recursions

$$
\begin{aligned}
q_t^* &= \Lambda_t^\circ q_{t+1}^*, \\
q_t^\circ &= \Phi_j' H_j' (F_j^\circ)^{-1} e_j + \Lambda_t^\circ q_{t+1}^\circ.
\end{aligned} \tag{167}
$$

The two sequences are incorporated into the smoothed estimates from $t = k-1$ down to $t = 1$ by the formula

$$
b_{t|T} = b_t + P_t^* q_{t+1}^* + P_t^\circ q_{t+1}^\circ. \tag{168}
$$

Conclusion

The Kalman filter is a complex device of great power and flexibility. Its exposition tends to generate an inordinate quantity of algebra. In the hands of the econometricians, the filter has undergone further developments that are conveyed in a literature that is challenging at the best of times.

One may expect that, when these developments are eventually assimilated into the mainstream of econometric methodology, some of their algebraic elaborations will fall into abeyance. Then a judgment will have been reached on which of the various competing formulations are the most useful or the most intelligible.

References

[1] Anderson, B.D.O., and J.B. Moore, (1979), *Optimal Filtering,* Prentice–Hall, Englewood Cliffs, New Jersey.

[2] Ansley, C.F., and R. Kohn, (1982), A Geometrical Derivation of the Fixed Interval Smoothing Equations, *Biometrika,* 69, 486–487.

[3] Ansley, C.F., and R. Kohn, (1985a), Estimation, Filtering and Smoothing in State Space Models with Incompletely Specified Initial Conditions, *The Annals of Statistics,* 13, 1286–1316.

[4] Ansley, C.F., and R. Kohn, (1985b), A Structured State Space Approach to Computing the Likelihood of an ARIMA Process and its Derivatives, *Journal of Statistical Computation and Simulation,* 21, 135–169.

[5] Ansley, C.F., and R. Kohn, (1990), Filtering and Smoothing in State Space Models with Partially Diffuse Initial Conditions, *Journal of Time Series Analysis,* 11, 275–293.

[6] Åström, K.J., U. Borisson, L. Ljung and B. Wittenmark, (1977), Theory and Applications of Self-Tuning Regulators, *Automatica,* 13, 457–476.

[7] Banerjee, A., R.L. Lumsdaine and J.H. Stock, (1992), Recursive and Sequential Tests of the Unit-Root Hypothesis: Theory and International Evidence, *Journal of Business and Economic Statistics,* 10, 271–287.

[8] Bell, W., (1984), Signal Extraction for Nonstationary Time Series, *The Annals of Statistics,* 12, 646–664.

[9] Bell, W., and S. Hillmer, (1991), Initialising the Kalman Filter for Nonstationary Time Series Models, *Journal of Time Series Analysis,* 12, 283–300.

[10] Bertrand, J., (1855), *Méthode des Moindres Carrés: Mémoires sur la combinaison des Observations par C-F. Gauss,* translation into French of *Theoria combinationis observationum erroribus minimis obnoxiae, by K.–F. Gauss,* Mallet-Bachelier, Paris.

[11] Bhargava, A., (1986), On the Theory of Testing for Unit Roots in Observed Time Series, *Review of Economic Studies,* 53, 359–384.

[12] Bomhoff, E.J., (1994), *Financial Forecasting for Business and Economics*, The Dryden Press, London.

[13] Box, G.E.P., and G.M. Jenkins, (1976), *Time Series Analysis: Forecasting and Control, Revised Edition,* Holden Day, San Francisco.

[14] Brown, R.L., J. Durbin and J.M. Evans, (1975), Techniques for Testing the Constancy of Regression Relationships over Time, *Journal of the Royal Statistical Society, Series B,* 37, 149–163.

[15] Burman, J.P., (1980), Seasonal Adjustment by Signal Extraction, *Journal of the Royal Statistical Society, Series A,* 143, 321–337.

[16] Canetti, R., and M.D. España, (1989), Convergence Analysis of the Least-Squares Identification Algorithm with a Variable Forgetting Factor for Time Varying Linear Systems, *Automatica,* 25, 609–612.

[17] Canova, F., and B.E. Hansen, (1995), Are Seasonal Patterns Constant over Time? A test for Seasonal Stability, *Journal of Business and Economic Statistics,* 13, 237–384.

[18] Chu, C-S.J, K. Hornik and C-M. Kuan (1995), The Moving-Estimates Test for Parameter Stability, *Econometric Theory,* 11, 699–720.

[19] Cleveland, W.P., and G.C. Tiao, (1976), Decomposition of Seasonal Time Series: A Model for the X-11 Program, *Journal of the American Statistical Association,* 71, 581–587.

[20] de Jong, P., (1988a), The Likelihood for a State Space Model, *Biometrika,* 75, 165–169.

[21] de Jong, P., (1988b), A Cross Validation Filter for Time Series Models, *Biometrika,* 75, 594–600.

[22] de Jong, P., (1989), Smoothing and Interpolation with the State Space Model, *Journal of the American Statistical Association,* 84, 1085–1088.

[23] de Jong, P., (1991a), The Diffuse Kalman Filter, *The Annals of Statistics,* 19, 1073–1083.

[24] de Jong, P., (1991b), Stable Algorithms for State Space Model, *Journal of Time Series Analysis,* 12, 143–157.

[25] de Jong, P., and SingFat Chu-Chun-Lin, (1994), Fast Likelihood Evaluation and Prediction for Nonstationary State Space Models, *Biometrika,* 81, 133–142.

[26] de Jong, P., and SingFat Chu-Chun-Lin, (2003), Smoothing with an Unknown Initial Condition, *The Journal of Time Series Analysis,* 24, 141–148.

[27] Diebold, F.X., (1986a), The Exact Initial Covariance Matrix of the State Vector of a General MA(q) Process, *Economic Letters,* 22, 27–31.

[28] Dickey, D.A. and W.A. Fuller, (1979), Distribution of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association,* 74, 427–431.

[29] Dickey, D.A., H.P. Hasza and W.A. Fuller, (1984), Testing for Unit Roots in Seasonal Time Series, *Journal of the American Statistical Association,* 79, 355–367.

[30] Diebold, F.X., (1986b), Exact Maximum-Likelihood Estimation of Autoregressive Models via the Kalman Filter, *Economic Letters,* 22, 197–201.

[31] Dufour, J-M., (1982), Recursive Stability Analysis of Linear Regression Coefficients, *Journal of Econometrics*, 19, 31–76.

[32] 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.

[33] Durbin, J., (1971), Boundary-Crossing Probabilities for the Brownian Motions and Poisson Processes and Techniques for Computing the Power of the Kolmogorov–Smirnov Test, *Journal of Applied Probability*, 8, 431–453.

[34] Durbin, J., and S.J. Koopman, (2001), *Time Series Analysis by State Space Methods,* Oxford University Press.

[35] Fortescue, T.R., L.S. Kershenbaum and B.E. Ydstie, (1981), Implementation of Self-Tuning Regulators with Variable Forgetting Factors, *Automatica,* 17, 831–835.

[36] Gardner, G., A.C. Harvey and G.D.A. Phillips, (1980), An Algorithm for Exact Maximum Likelihood Estimation of Autoregressive Moving Average Models by Means of Kalman Filtering, Algorithm AS 154, *Applied Statistics,* 29, 311–322.

[37] Gauss, K.F., 1777–1855, (1809), *Theoria Motus Corporum Celestium, English translation by C.H. Davis (1857).* Reprinted 1963, Dover Publications, New York.

[38] Gauss, K.F., 1777–1855, (1821, 1823, 1826), *Theoria combinationis observationum erroribus minimis obnoxiae, (Theory of the combination of observations least subject to error),* French translation by J. Bertrand (1855), *Méthode de Moindres Carrés: Mémoires sur la combinaison des Observations par C.–F. Gauss,* Mallet–Bachelier, Paris, English translation by G.W. Stewart (1995), Classics in Applied Mathematics no. 11, SIAM Press, Philadelphia.

[39] Gersch, W., and G. Kitigawa, (1983), Prediction of Time Series with Trends and Seasonalities, *Journal of Business and Economic Statistics,* 1, 253–256.

[40] Goméz, V., and A. Maravall, (1994), Initialising the Kalman Filter with Incompletely Specified Initial Conditions, pages 39–62 in Guanring Chen (ed.) *Approximate Kalman Filtering,* World Scientific Publishing Co., Singapore.

[41] Harrison, P.J., and C.F. Stevens, (1976), Bayesian Forecasting (With a Discussion), *Journal of the Royal Statistical Society, Series B,* 38, 205–247.

[42] Harvey, A.C., (1989), *Forecasting, Structural Time Series Models and the Kalman Filter,* Cambridge University Press, Cambridge.

[43] Harvey, A.C., (1990), *The Econometric Analysis of Time Series: Second Edition,* Philip Allan, London.

[44] Harvey, A.C. and P. Collier, (1977), Testing for Functional Misspecification in Regression Analysis, *Journal of Econometrics,* 6, 103–119.

[45] Harvey, A.C., and P. Todd, (1983), Forecasting Economic Time Series with Structural and Box–Jenkins Models: A Case Study, *Journal of Business and Economic Statistics*, 1, 299–307.

[46] Hillmer, S.C., and G.C. Tiao, (1982), An ARIMA-Model-Based Approach to Seasonal Adjustment, *Journal of the American Statistical Association,* 77, 63–70.

[47] Hylleberg, S., R.F. Engle, C.W.J. Granger and B.S. Yoo, (1990), Seasonal Inte-

gration and Co-integration, *Journal of Econometrics,* 44, 215–238, reprinted in S. Hylleberg (ed.), *Modelling Seasonality,* Oxford University Press, Oxford.

[48] Jones, R., (1980), Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. *Technometrics,* 22, 389–395.

[49] Kalman, R.E., (1960), A New Approach to Linear Filtering and Prediction Problems, *Transactions of the ASME Journal of Basic Engineering,* Series D, 82, 35–45.

[50] Kalman, R.E., and R.S. Bucy, (1961), New Results in Linear Filtering and Prediction Theory, *Transactions of the ASME Journal of Basic Engineering,* Series D, 83, 95–107.

[51] Kiparissides, C., and S.L. Shah, (1983), Self-Tuning and Stable Adaptive Control of a Batch Polymerisation Reactor, *Automatica,* 19, 225–235.

[52] Kohn, R., and C.F. Ansley, (1986), Estimation, Prediction and Interpolation for ARIMA Models with Missing Data, *Journal of the American Statistical Association,* 81, 751–761.

[53] Kohn, R., and C.F. Ansley, (1987), Signal Extraction for Finite Nonstationary Time Series, *Biometrika,* 74, 411–421.

[54] Kohn, R., and C.F. Ansley, (1989), A Fast Algorithm for Signal Extraction, Influence and Cross-Validation in State Space Models, *Biometrika,* 76, 65–79.

[55] Koopman, S.J., (1993), Disturbance Smoother for State Space Models, *Biometrika,* 80, 117–126.

[56] Koopman, S.J., (1997), Exact Initial Kalman Filtering and Smoothing for Nonstationary Time Series Models, *Journal of the American Statistical Association,* 92, 1630–1638.

[57] Koopman, S.J., N. Shephard and J.A. Doornik, (1999), Statistical Algorithms for Models in State Space using SsfPack 2.2, *Econometrics Journal,* 2, 107–160.

[58] Krämer, W., W. Ploberger, and R. Alt, (1988), Testing for Structural Change in Dynamic Models, *Econometrica,* 56, 1355-1369.

[59] Kuan, C-M., (1998), Tests for Changes in Models with a Polynomial Trend, *Journal of Econometrics,* 84, 75–91.

[60] Kuan, C-M., and K. Hornik (1995), The Generalised Fluctuation Test: A Unifying View, *Econometric Reviews,* 14, 135–161.

[61] Kwiatkowski, D., P.C.B. Phillips, P. Schmidt and Y. Shin, (1992), Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root, *Journal of Econometrics,* 54, 159–178.

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

[63] Lozano, R., (1983), Convergence Analysis of Recursive Identification Algorithms with Forgetting Factors, *Automatica,* 19, 95–97.

[64] Maravall, A., (1985), On Structural Time Series Models and the Characterisation of Components, *Journal of Business and Economic Statistics*, 3, 350–355.

[65] Meditch, J.S., (1973), A Survey of Data Smoothing for Linear and Nonlinear Dynamic Systems, *Automatica,* 9, 151–162.

[66] Mélard, G., (1983), A Fast Algorithm for the Exact Likelihood of Autoregressive Moving Average Time Series, Algorithm AS 197, *Applied Statistics,* 32, 104–114.

[67] Merkus, H.R., D.S.G. Pollock and A.F. de Vos, (1993), A Synopsis of the Smoothing Formulae Associated with the Kalman Filter, *Computational Economics,* 6, 177–200.

[68] Mittnik, S., (1987a), The Determination of the State Covariance Matrix of Moving-Average Processes without Computation, *Economic Letters,* 23, 177–179.

[69] Mittnik, S., (1987b), Non-Recursive Methods for Computing The Coefficients of the Autoregressive and Moving-Average Representation of Mixed ARMA Processes, *Economic Letters,* 23, 279–284.

[70] Phillips, P.C.B., (1987), Time Series Regressions with a Unit Root, *Econometrica,* 55, 277–301.

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

[72] Ploberger, W., W Krämer and K. Kontros, (1989), A New Test for Structural Stability in the Linear Regression Model, *Journal of Econometrics,* 40, 307–318.

[73] Pollock, D.S.G., (1979), *The Algebra of Econometrics,* John Wiley and Sons, Chichester.

[74] Pollock, D.S.G., (1999), *Time-Series Analysis, Signal Processing and Dynamics,* Academic Press, London.

[75] Pollock, D.S.G., (2000), Trend Estimation and De-trending via Rational Square Wave Filters, *Journal of Econometrics,* 99, 317–334.

[76] Pollock, D.S.G., (2001), Filters for Short Non-stationary Sequences, *Journal of Forecasting,* 20, 341–355.

[77] Pollock, D.S.G., (2001), The Methodology for Trend Estimation, *Economic Modelling,* 18, 75–96.

[78] Pollock, D.S.G., (2003), Improved Frequency-Selective Filters, *Computational Statistics and Data Analysis,* 42, 279–297.

[79] Rauch, H.E., (1963), Solutions to the Linear Smoothing Problem, *IEEE Transactions on Automatic Control,* AC-8, 371–372.

[80] Rosenberg, B., (1973), Random Coefficient Models: The Analysis of a Cross Section of Time Series by Stochastically Convergent Parameter Regression, *Annals of Economics and Social Measurement,* 2, 399–428.

[81] Sanoff, S.P., and P.E. Wellstead, (1983), Comments on: 'Implementation of Self-Tuning Regulators with Variable Forgetting Factors', *Automatica,* 19, 345–346.

[82] Schweppe, F.C., (1965), Evaluation of Likelihood Functions for Gaussian Signals, *IEEE Transactions on Information Theory,* 11, 61–70.

[83] Smith, R.J., and A.M.R. Taylor, (2001), Recursive and Rolling Regression-based Tests of the Seasonal Unit Root Hypothesis, *Journal of Econometrics,*

105, 309–336.
[84] Snyder, R.D., (1988), Computational Aspects of Kalman Filtering with a Diffuse Prior Distribution, *Journal of Statistical Computation and Simulation,* 29, 77–86.
[85] Stigler, S.M., (1986), *The History of Statistics,* Harvard University Press, Cambridge, Mass.
[86] Stock, J.H., (1994), Unit Roots, Structural Breaks and Trends, Chapter 46 in *Handbook of Econometrics, Volume IV,* Elsevier Science, Amsterdam.
[87] Theil, H., and A.S. Goldberger, (1961), On Pure and Mixed Statistical Estimation in Economics, *International Economic Review,* 2, 65–78.
[88] Theil, H., (1963), On the Use of Incomplete Prior Information in Regression Analysis, *Journal of the American Statistical Association,* 58, 401–414.
[89] Theil, H., (1971), *Principles of Econometrics,* John Wiley and Sons, New York.
[90] Wellstead, P.E., and M.B. Zarrop, (1991), *Self-tuning Systems: Control and Signal Processing,* John Wiley and Sons, Chichester.
[91] Young, P., (1984), *Recursive Estimation and Time-Series Analysis,* Springer Verlag, Berlin.
[92] Zarrop, M.B., (1983), Variable Forgetting Factors in Parameter Estimation, *Automatica,* 19, 295–298.

A Appendix

The Partitioned Matrix Inverse: If $A = A'$ and $C = C'$ are full rank symmetric matrices, then

$$\begin{bmatrix} A & B \\ B' & C \end{bmatrix} = \begin{bmatrix} I & BC^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BC^{-1}B' & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} I & 0 \\ C^{-1}B' & I \end{bmatrix}, \tag{A.1}$$

whence

$$\begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ 0 & -C^{-1}B' \end{bmatrix} \begin{bmatrix} (A - BC^{-1}B')^{-1} & 0 \\ 0 & C^{-1} \end{bmatrix} \begin{bmatrix} I & -BC^{-1} \\ 0 & I \end{bmatrix} \tag{A.2}$$

$$= \begin{bmatrix} (A - BC^{-1}B')^{-1} & -(A - BC^{-1}B')^{-1}BC^{-1} \\ -C^{-1}B'(A - BC^{-1}B')^{-1} & C^{-1} + C^{-1}B'(A - BC^{-1}B')^{-1}BC^{-1} \end{bmatrix}.$$

These results are confirmed by direct multiplication.

The Matrix Inversion Lemma: In reference to (A.2), there are the following matrix identities:

$$\text{(i)} \quad (C - B'A^{-1}B)^{-1} = C^{-1} + C^{-1}B'(A - BC^{-1}B')^{-1}BC^{-1}, \tag{A.3}$$

$$\text{(ii)} \quad (A - BC^{-1}B')^{-1} = A^{-1} + A^{-1}B(C - B'A^{-1}B)^{-1}B'A^{-1},$$
$$\text{(iii)} \quad (C + B'A^{-1}B)^{-1} = C^{-1} - C^{-1}B'(A + BC^{-1}B')^{-1}BC^{-1}.$$

Results (i) and (ii) are proved by comparing

$$\begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & (C - B'A^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -B'A^{-1} & I \end{bmatrix} \tag{A.4}$$
$$= \begin{bmatrix} A^{-1} + A^{-1}B(C - B'A^{-1}B)^{-1}B'A^{-1} & -A^{-1}B(C - B'A^{-1}B)^{-1} \\ -(C - B'A^{-1}B)B'A^{-1} & (C - B'A^{-1}B)^{-1} \end{bmatrix}$$

with (A.2) above. To prove (iii), C is replaced in (i) by $-C$ and both sides of the equation are multiplied by -1.

The Partitioned Normal Distribution: The probability density function of a normal vector x of n elements with a mean vector of $E(x) = \mu$ and a dispersion matrix of $D(x) = \Sigma$ is

$$N(x; \mu, \Sigma) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp[-\{x - E(x)\}'\Sigma^{-1}\{x - E(x)\}/2]. \tag{A.5}$$

If $x = [x_1', x_2']'$, then the quadratic function $S(x) = \{x - E(x)\}'\Sigma^{-1}\{x - E(x)\}$ may be partitioned conformably to give

$$S(x_1, x_2) = \begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2) \end{bmatrix}' \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2) \end{bmatrix} \tag{A.6}$$
$$= \begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2|x_1) \end{bmatrix}' \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2|x_1) \end{bmatrix}$$
$$= \begin{bmatrix} x_1 - E(x_1|x_2) \\ x_2 - E(x_2) \end{bmatrix}' \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - E(x_1|x_2) \\ x_2 - E(x_2) \end{bmatrix},$$

where

$$\begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2|x_1) \end{bmatrix} = \begin{bmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2) \end{bmatrix}, \tag{A.7}$$
$$\begin{bmatrix} x_1 - E(x_1|x_2) \\ x_2 - E(x_2) \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 - E(x_1) \\ x_2 - E(x_2) \end{bmatrix}.$$

These results follow immediately from (A.2) and (A.4).

The Calculus of Conditional Expectations: Consider the jointly distributed normal random vectors x and y which bear the linear relationship $E(y|x) = \alpha + B'\{x - E(x)\}$. Then, the following conditions apply:

$$\begin{aligned}
&\text{(i)} && E(y|x) = E(y) + C(y,x)D^{-1}(x)\{x - E(x)\}, \\
&\text{(ii)} && D(y|x) = D(y) - C(y,x)D^{-1}(x)C(x,y), \\
&\text{(iii)} && E\{E(y|x)\} = E(y), \\
&\text{(iv)} && D\{E(y|x)\} = C(y,x)D^{-1}(x)C(x,y), \\
&\text{(v)} && D(y) = D(y|x) + D\{E(y|x)\}, \\
&\text{(vi)} && C\{y - E(y|x), x\} = 0.
\end{aligned} \tag{A.8}$$

These results are obtained from (A.6) and (A.7) by setting $x_1 = y$, $x_2 = x$, $\Sigma_{11} = D(y)$, $\Sigma_{22} = D(x)$ and $\Sigma_{12} = C(y,x)$. Then, it is recognised that $\alpha = E(y)$ and $B' = C(y,x)D^{-1}(x) = \Sigma_{12}\Sigma_{22}^{-1}$.