Department of Economics

Recursive Estimation in Econometrics

Stephen Pollock

Working Paper No. 462 June 2002 ISSN 1473-0278

Queen Mary
University of London

RECURSIVE ESTIMATION IN ECONOMETRICS

by D.S.G. Pollock

Queen Mary, University of London

Email: stephen_pollock@sigmapi.u-net.com

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

Date: May, 2002

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

JEL Classification: C22

1. Introduction

This paper is a tutorial review of recursive estimation which originates in the author's own needs to understand, to use and, occasionally, to amend or to supplant the algorithms in question.

The algorithms of recursive estimation and of Kalman filtering have been used increasingly in applied econometrics in the past two decades, albeit that econometricians have been slower in exploiting them that have other statisticians. Reasons for this tardiness are suggested in the next section of the paper which deals with some historical aspects of recursive estimation.

The third section of the paper lays some essential groundwork by expounding the algorithm of ordinary recursive regression. This can be seen as a preparation for the complexities of the Kalman filter, the features of which can be more easily understood if they can be related to something simpler which has the same architecture.

The treatment of recursive regression, in section 3, has a Bayesian flavour, and it relies heavily upon the calculus of conditional expectations, of which the essential results are provided in an appendix. Section 4 deals with the prediction-error decompositions associated with recursive regression, whilst Section 5, which deals with extensions and elaborations of recursive regression, mentions some applications in control engineering which could be exploited by econometricians.

Section 6, embarks on a treatment of the Kalman filter which is depicted as an elaboration of the regression algorithm in a manner which reflects the preceding derivations. The three succeeding sections, which deal with the likelihood function and the starting-value problem, benefit from the treatment of the analogous problems in the regression context; and the treatment of this problem has consequences for the smoothing operations described in section 10.

An extensive bibliography is also provided which contains references to some of the work of the econometricians on the problems of recursive estimation together with some of the sources on which they have relied. Because of the complexity and diversity of the notation, readers of this material might be advised to maintain a glossary to assist them is making the necessary translations and comparisons.

Many of the contributions to the literature on Kalman filtering assume a considerable familiarity with the associated algebra. Some of the principal contributions of the econometricians have come is small increments conveyed in long sequences of papers. These papers are never entirely self-contained and, often, they refer only to their immediate predecessors. Seldom do they recapitulate the original motivations. The task of collating such literature makes for difficult reading. One of the purposes of present paper is to gather the important results and the ideas that lie behind them within a small compass.

2. Historical Aspects

The concept of least-squares regression originates with two people. It is nowadays accepted that Legendre (1805) was responsible for the first published account of the theory; and it was he who coined the term *Moindes Carrés* or least squares. However, it was Gauss who developed the method as a statistical tool by embedding it in a context which involved a probabilistic treatment of errors of observation. Confusion over the rival claims of priority arises from the fact that, although his first published exposition of the method appeared in 1809 in *Theoria Motus Corporum Celestium*, when he was 31 years of age, Gauss claimed that he had formulated his ideas many years earlier when he was in his early twenties. These matters are dealt with in the book of Stigler (1986) on the History of Statistics.

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

The relevance of Gauss's second exposition to the theory of recursive least-squares estimation and to the concept of the Kalman filter lies in a brief passage where Gauss shows that it is possible *to find the changes which the most likely values of the unknowns undergo when a new equation is adjoined, and to determine the weights of these new determinations.* This passage refers to the business of augmenting the normal equations when a new observation becomes available. In effect, Gauss developed the algorithm of recursive least-squares estimation. The French translation of the passage in question, which is due to Bertrand (1855), has been reproduced by Young (1984) in an appendix of his book, where it is accompanied by a synoptic commentary which 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 on two separate occasions. The first rediscovery by Plackett (1950) was before the advent of efficient on-line electronic computing; and this also passed almost unnoticed. It was the second rediscovery of the recursive algorithms in 1960 in the context of control theory which was the cue to a rapid growth of interest. Stemming from the papers of Kalman (1960) and Kalman and Bucy (1961) a vast literature on Kalman filtering has since accumulated.

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

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

Since Kalman's seminal paper, several other derivations have been offered, and a welter of alternative notation has arisen. Most of the alternative derivations attempt to avoid the concepts of Hilbert space and to reduce the terminology of the derivation to something closer to that of the ordinary theory of least-squares regression. Other derivations have been from a maximum-likelihood or a Bayesian standpoint.

The derivation which, at first, attracted the attention of econometricians is that of Duncan and Horn (1972). This exploits the concept of mixed estimation which originated with Theil and Goldberger (1961) and which was extended by Theil (1963). An account of the method is to be found in the textbook of Theil (1971, pps. 347–352). More recently, there has been a tendency to adopt a Bayesian approach, as in the recent book of Durbin and Koopman (2001), for example.

Econometricians have been slow to adopt the Kalman filter, partly because they have been reluctant to espouse the notion of time-varying parameters. They have tended to adhere to notions of parametric constancy and to imagine that their structural models will break at identifiable points rather than flex or bend.

The principal use of the Kalman filter by econometricians, together with the associated fixed-interval smoothing algorithms, has been in trend estimation and signal extraction, of which there is now a considerable literature. The work of Harrison and Stevens (1976), which foreshadowed the development of structural time series models, has been highly influential in this connection as have the articles of Harvey and Todd (1983) and Gersch and Kitagawa (1983) and the book of Harvey (1989).

An equal influence in favour of an alternative methodology, which has tended to be implemented by means other than the Kalman filter, such as Burman's (1980) method, has been exerted by Cleveland and Tiao (1976), Hillmer and Tiao (1982) and by Maravall (1985). Much of the relevant literature has been cited in the author's own recent contributions to the area—see Pollock (2000, 2001a, 2001b, 2002)—which also employ alternatives to the Kalman filter.

Another use of the Kalman filter that has been increasing in recent years is as a device for calculating the likelihood function of a time series model for the purpose of estimating its parameters. A requirement is that the model should be represented in state-space form, whereafter the likelihood function can be evaluated via the prediction-error decomposition in the manner that was originally indicated by Schweppe (1965).

Early examples from econometrics were the algorithms for evaluating the likelihood of autoregressive moving-average (ARMA) models that were published by Gardner Harvey and Phillips (1980) and by Mélard (1983). Jones (1980) used this approach in fitting ARMA models to time series with missing observations. A variety of state-space representations for ARMA models have been described by Pollock (1999). However, the applications of this method of evaluating the likelihood function extend, nowadays, far beyond the classical univariate time series models.

Symptomatic of the growing use by econometricians of the Kalman filter and of other recursive algorithms is the availability of accessible software which they have originated. Examples are the *SsfPack* software which has been described by Koopman, Shephard and Doornick (1999) and the software that has been provided by Bomhoff (1994) in association with his book.

The scientific community as a whole is well served nowadays by freely available resources relating to the Kalman filter; and 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 results in the algebra of conditional expectations presented in the appendix to derive the algorithm for the recursive estimation of the parameters of a classical linear regression model. The tth instance of the regression relationship is represented by

$$(1) \qquad y_t = x_t'\beta + \varepsilon_t,$$

where y_t is a scalar value and x_t has k elements. It is assumed that the disturbances ε_t are serially independent and normally distributed with

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

In order to initiate the recursion, there must be an initial estimate of β together with a corresponding dispersion matrix. In the usual context of classical regression theory, we should regard this dispersion matrix as the variance–covariance matrix of the estimator. Instead, we are inclined to attribute a distribution to β and to regard $b_0 = E(\beta)$ and $\sigma^2 P_0 = D(\beta)$ as its mean and its dispersion matrix. This distribution is, in effect, a Bayesian prior.

The information $\mathcal{I}_t$, available at time t, is the set of observations together with $\mathcal{I}_0$, which is the set $\{\beta_0, \sigma^2 P_0\}$ if there is prior information and which is the emptyset in the absence of such information. Thus, $\mathcal{I}_t = \{y_t, \mathcal{I}_{t-1}\} = \{y_t, \ldots, y_1, \mathcal{I}_0\}$. We shall work, initially, under the presumption that the prior distribution of β is fully specified, in which case it gives rise of 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 in a manner which makes best use of the previous estimates $b_{t-1} = E(\beta|\mathcal{I}_{t-1})$ and $\sigma^2 P_{t-1} = D(\beta|\mathcal{I}_{t-1})$. The first task is to evaluate the expression

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

which is derived directly from (A.8.i). There are three elements on the RHS which require further development. The first is the term

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

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

According to the result (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 is a function solely of the information in $\mathcal{I}_{t-1} = \{y_{t-1}, \mathcal{I}_{t-2}\}$. By pursuing this argument back to the start of the sample, it can be established that the prediction errors 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 $\mathcal{I}_t = \{h_t, \ldots, h_1, \mathcal{I}_0\}$.

Next is the dispersion matrix associated with the prediction. This is

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

Finally, there is the covariance

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

On putting these elements together, 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 of 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 this is

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

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

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

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

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

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

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

Alternative expressions are available for P_t and κ_t:

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

(16) $$\kappa_t = P_t x_t.$$

The expression on the RHS of (15) is confirmed by using the matrix inversion formula given by (A.3.iii) to recover the original expression for P_t given under (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 $P_t^{-1}P_{t-1}x_t = x_t f_t$. The latter is readily confirmed using the expression for P_t from (15) and the expression for f_t from (11).

Equation (15) indicates that

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

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

When equations (15) and (16) are used in (13), we get the following expression for recursive least-squares estimate:

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

The formula of (18) certainly appears to be simpler than that of (7). However, in comparison to the latter, it is computationally inefficient. The formula of (7) entails finding the inverse of the scalar element $f_t = x_t P_{t-1} x_t' + 1$, which is the factor in the dispersion of the prediction error. The formula under (18) involves the inversion of the entire matrix P_t. To use this formula in place of that of (7) would be to loose all the computational advantages of the recursive least-squares algorithm.

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

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

By pursuing a recursion on the RHS and using the expression from (17) on the LHS, it is found that $\big(P_0^{-1} + \sum_{i=1}^{t} x_i x_i'\big) b_T = P_0^{-1} b_0 + \sum_{i=1}^{t} x_i y_i$. 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.

In reality, whenever the formulae are used in pursuit of an ordinary regression analysis, the initial estimate of the parameter vector and the corresponding dispersion matrix are liable to be determined by an initial stretch of data. Thus, if $X_k = [x_1, \ldots, x_k]'$ denotes a full-rank matrix of k initial observations of the regressors and if $Y_k = [y_1, \ldots, y_k]'$ denotes the corresponding vector of observations of the dependent variable, then the recursion starts with $b_k = X_k^{-1}Y_k$ and $P_k = (X_k'X_k)^{-1}$. Moreover, the full-sample estimator becomes the ordinary least-squares estimator $b = (X'X)^{-1}X'y$.

To understand the status of the initial solution b_k, one should think of an arbitrarily chosen finite value of b_0 together with a dispersion matrix P_0 containing very large diagonal elements to reflect the 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 $b_1, \ldots, b_k$, one should discover that b_k is within a epsilon of the value given by $X_k^{-1}Y_k$

There are other, more precise, ways of initialising the recursive procedure which use pseudo information, or ‘diffuse’ information, to enable the iterations to begin at $t = 0$. By the time $t = k + 1$, when there is sufficient empirical information to determine a unique parameter estimate, the system should be purged of the pseudo information.

To describe such a method, let us resolve the dispersion matrix of the estimated state into two components such that $P_t = P_t^* + \rho P_t^\circ$, where P_t^* relates to the sample information and where P_t° relates to the diffuse presample information. The latter is intended only for the purpose of initialising the filter at time $t = 0$. As the observations accrue, we should seek to incorporate the new information into P_t^* and to remove from P_t° any pseudo information that might conflict with it.

In order to implement the updating formulae, it is necessary to find expressions for f_t^{-1} and κ_t which reflect the nature of the available information. Therefore, let

$$f_t^* = x_t'P_t^*x_t + 1, \qquad f_t^\circ = x_t'P_t^\circ x_t \qquad \text{and} \qquad f_t = f_t^* + \rho f_t^\circ. \tag{22}$$

Then $f_t = \rho f_t^\circ(1 - \rho^{-1}q)$, where $q = -f_t^*/f_t^\circ$; and, since $\rho > 1$, it follows that there is a series expansion of the inverse in the form of

$$\begin{aligned} f_t^{-1} &= \frac{1}{\rho f_t^\circ}\left(1 + \frac{q}{\rho} + \frac{q^2}{\rho^2} + \cdots\right) \\ &= \frac{g_1}{\rho} + \frac{g_2}{\rho^2} + \frac{g_3}{\rho^3} + \cdots. \end{aligned} \tag{23}$$

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 whilst 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 will 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

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

With $\rho \to \infty$, only the first term of (26) survives in isolation; which gives $\kappa_t = P_{t-1}^\circ x_t (f_t^\circ)^{-1}$. Therefore, 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, we consider the updating equation for the dispersion of the estimate. This embodies

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

By carrying the leading terms of this expression into equation (14) and separating $P_t = P_t^* + \rho P_t^\circ$ into its two parts, we obtain following equations:

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

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

The updating equation of (30), which is associated with the diffuse information, has the form of $P_t^\circ = (I - Q)P_{t-1}^\circ$, where $Q = P_{t-1}^\circ x_t (x_t' P_{t-1}^\circ x_t)^{-1} x_t'$ is an idempotent matrix such that $Q = Q^2$ and $I - Q = (I - Q)^2$. Thus, P_t° is formed by projecting P_{t-1}° onto the subspace which is orthogonal to x_t. Unless $P_{t-1}^\circ x_t = 0$, which is unlikely when $P_t^\circ \neq 0$, the matrix $I - Q$ will have less than full rank; and, therefore, $\mathrm{Rank}(P_t^\circ) < \mathrm{Rank}(P_{t-1}^\circ)$.

Eventually, the loss of rank will lead to $P_t^\circ = 0$. From that point on, 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 specified under (10)–(14). In the absence informative prior information, the procedure can be initialised with $P_0^* = 0$, $P_0^\circ = I$ and $b_0 = 0$. After k steps, it is to be expected that the observation vectors $x_1', \ldots, x_k'$ will fill the k-dimensional space in which the parameter estimates reside; and this will be the point at which the transition to the standard recursions occurs.

The algorithm that we have described was proposed by Ansley and Kohn (1985a) who developed it in the context of the Kalman filter, where it has a more significant role to play. The essential features of the exposition above are due to Koopman (1997) and to Durbin and Koopman (2001).

4. 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, there is $E(\beta) = b_0$ and $D(\beta) = P_0$. Combining these elements gives

$$\begin{aligned} E(y) &= XE(\beta) + E(\varepsilon) \\ &= Xb_0 \end{aligned} \quad \text{and} \quad \begin{aligned} D(y) &= XD(\beta)X' + D(\varepsilon) \\ &= XP_0X' + \sigma^2 I. \end{aligned} \tag{32}$$

The marginal density function of y is

$$N(y) = (2\pi\sigma)^{-T/2} |XP_0X' + I|^{-1/2} \exp\{-S(y)/(2\sigma^2)\}, \tag{33}$$

of which the quadratic exponent is

$$\begin{aligned} S(y) &= (y - Xb_0)'(XP_0X' + I)^{-1}(y - Xb_0) \\ &= (y - Xb_0)'\{I - X(X'X + P_0^{-1})^{-1}X'\}(y - Xb_0). \end{aligned} \tag{34}$$

Here, the second equality follows from the matrix identity of (A.3.iii).

The recursive regression algorithm, which is described by equations (10)–(14), entails a decomposition of the marginal function $N(y)$, described as the prediction-error decomposition. This takes the form of

$$N(y_1,\ldots,y_T;\mathcal{I}_0) = N(y_1;\mathcal{I}_0)\prod_{t=2}^{T} N(y_t|\mathcal{I}_{t-1}). \tag{35}$$

The explicit form of one of the factors of this decomposition, when $t > 1$, is

$$N(y_t|\mathcal{I}_{t-1}) = (2\pi\sigma^2 f_t)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\frac{(y_t - x_t'b_{t-1})^2}{1 + x_t'P_{t-1}x_t}\right\}. \tag{36}$$

The marginal density function $N(y_1;\mathcal{I}_0)$, which is the first factor of the decomposition, is obtained by specialising the expression of (33) for $N(y)$ to the case of a single observation; and this is also obtained from $N(y_t|\mathcal{I}_{t-1})$ by setting $t = 1$. It follows that the quadratic function of (34) can be written alternatively as

$$S(y) = \sum_{t=1}^{T}\frac{(y_t - x_t'b_{t-1})^2}{1 + x_t'P_{t-1}x_t} = \sum_{t=1}^{T}\frac{h_t^2}{f_t} = \sum_{t=1}^{T} w_t^2. \tag{37}$$

It can be demonstrated that, given the true values of the parameters, there is a one-to-one correspondence between the errors $\xi_t = y_t - x_t'b_0$ and the recursive residuals $h_t = y_t - x_t'b_{t-1}$. Consider the basic recursive formula:

$$\begin{aligned} b_t &= b_{t-1} + \kappa_t(y_t - x_t'b_{t-1}) \\ &= \lambda_t b_{t-1} + \kappa_t y_t, \end{aligned} \tag{38}$$

where $\lambda_t = 1 - \kappa_t x_t'$. By running this recursion from the start for a few iterations, we get

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2\lambda_1 \\ \lambda_3\lambda_2\lambda_1 \end{bmatrix} b_0 - \begin{bmatrix} \kappa_1 & 0 & 0 \\ \lambda_2\kappa_1 & \kappa_2 & 0 \\ \lambda_3\lambda_2\kappa_1 & \lambda_3\kappa_2 & \kappa_3 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{39}$$

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

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -x_2'\kappa_1 & 1 & 0 & 0 \\ -x_3'\lambda_2\kappa_1 & -x_3'\kappa_2 & 1 & 0 \\ -x_4'\lambda_3\lambda_2\kappa_1 & -x_4'\lambda_3\kappa_2 & -x_4'\kappa_3 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} - \begin{bmatrix} x_1' \\ x_2'\lambda_1 \\ x_3'\lambda_2\lambda_1 \\ x_4'\lambda_3\lambda_2\lambda_1 \end{bmatrix} b_0. \tag{40}$$

On defining $\lambda_{j,m} = \lambda_j\lambda_{j-1}\cdots\lambda_m$ with $\lambda_{j,j} = \lambda_j$ and $\lambda_{j,j+1} = 1$, the generic expression for the prediction error becomes

$$\begin{aligned} h_t &= y_t - x_t'b_{t-1} \\ &= y_t - x_t'\sum_{j=1}^{t-1}\lambda_{t-1,j+1}\kappa_j y_j - x_t'\lambda_{t-1,1}b_0. \end{aligned} \tag{41}$$

Equation (40) can be written in summary notation as $h = Ly - Wb_0$. But $E(h) = 0$ and $E(y) = Xb_0$; so the equations indicate that $LXb_0 = Wb_0$, which is to say that $W = LX$, since b_0 can take any value. Substituting this back into the original equation gives $h = L(y - Xb_0)$, which holds for any extension of the recursion. (The equality $W = LX$ can, of course, be demonstrated by purely algebraic means without resort to the expectations operator.) Thus, it follows that the marginal sum of squares of (34) can also be written as

$$\begin{aligned} S(y) &= (y - Xb_0)'(XP_0X' + I)^{-1}(y - Xb_0) \\ &= (y - Xb_0)'L'F^{-1}L(y - Xb_0) = h'F^{-1}h, \end{aligned} \tag{42}$$

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

The case where there is no prior information about β may be handled by concentrating the likelihood function $N(y)$ in respect of b_0 and P_0. It transpires that the minimising value is the ordinary least-squares estimator $b = (X'X)^{-1}X'y$. (A means of reaching this result will be demonstrated in section 7.) It is also evident that the minimising value of P_0 is zero.

According to the normal understanding, the condition that $P_0 = 0$ signifies that there is complete information regarding the value of β, with the effect that it becomes a know constant. This is clearly at variance with the actual circumstance that there is no prior information regarding β. The anomaly may be taken as an reflection of the fact that the appropriate criterion for deriving the estimate of β, in the absence of prior information, is the minimisation of 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

$$S^c(y) = e'e = y\{I - X(X'X)^{-1}X'\}y = \varepsilon'\{I - X(X'X)^{-1}X'\}\varepsilon, \tag{43}$$

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 retains a prediction-error decomposition which is in the form of (37), with the index of summation beginning at $t = k + 1$, instead of at $t = 1$. and with $b_k = X_k^{-1}Y_k$ and $P_k = (X_k'X_k)^{-1}$ as starting values (see, for example, Pollock 1999, p. 249). 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 between 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 which is analogous to that of (42):

$$\begin{aligned} S^c(y) &= (y_2 - X_2b_1)'\{X_2(X_1'X_1)^{-1}X_2' + I\}^{-1}(y_2 - X_2b_1) \\ &= (y_2 - X_2b_1)'L_2'F_2^{-1}L_2(y_2 - X_2b_1) = h_2'F_2^{-1}h_2. \end{aligned} \tag{44}$$

Here, the matrices L_2 and $F_2 = \text{diag}\{f_{k+1}, \ldots, f_T\}$ are also analogous to those defined in respect of of equation (42). The vector $h_2 = [h_{k+1}, \ldots, h_T]'$ contains the prediction errors, of which the normalised versions $w_t = h_t/f_t$ are in the vector w. The essential conditions affecting the recursive residuals are that

$$E(w) = 0 \quad \text{and} \quad D(w) = \sigma^2 I_{T-k}, \tag{45}$$

which is to say that they possess a spherical distribution.

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. Thus, as Harvey (1990) has indicated, the recursive residuals are amenable to an exact von Neumann ratio test aimed at detecting serial correlation in the disturbances. This can be used in preference to the Durbin–Watson test constructed from the ordinary least-squares residuals. Since the least-squares residuals are dependent on the values in the matrix X, it is not possible to derive exact significance points that apply to every instance of that test.

Another leading use of the recursive residuals is in the CUSUM test proposed by Brown, Durbin and Evans (1975) and in the various derivatives of this test. The test, which is aimed at detecting instability in the regression parameters, rejects the null hypothesis of parametric invariance if the trajectory of the cumulative sum of the recursive residuals crosses an upper or a 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 to the CUSUM process with increasing accuracy as the sample size increases—see Durbin (1971).

The CUSUM test has been further investigated by several authors including Dufour (1982) and Krämer, Ploberger and Alt, (1988). The latter have investigated the use of the CUSUM test when there are lagged dependent variables amongst the regressors; and they have shown that it retains its asymptotic significance level 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.

There is a variety of alternative residuals associated with the classical regression model which 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. It will be 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 transformation T such that $TX'XT' = I$ and $T'T = (X'X)^{-1}$. 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 matrix of order $T \times (T-k)$ which is complimentary to C_1 such 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

$$(46)\qquad \begin{aligned} S^c(y) &= \sum_{t=1}^{T} e_t^2 = y'\{I - X(X'X)^{-1}X'\}y \\ &= y'C_2C_2'y = \sum_{t=k+1}^{T} v_t^2; \end{aligned}$$

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

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

$$E(v) = 0 \quad \text{and} \quad D(v) = C_2' E(\varepsilon\varepsilon')C_2 = \sigma^2 I_{T-k}, \tag{47}$$

which shows that the LUS residuals also possess a spherical distribution.

5. Extensions of the Recursive Least-Squares Algorithm

The algorithm which we have presented in the previous sections represents little more than an alternative means of computing the ordinary least-squares regression estimates. If the parameters of the underlying process that generates the data are stable, then we can expect the estimate b_t to converge to a stable value also 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.

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

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

One way of accommodating parametric variability is to base the estimate on only the most recent portion of the data. As each new observation is acquired another observation may be removed so that, at any instant, the estimator comprises only n points. Such an estimator has been described as a rolling regression. Implementations are available in the recent versions of the more popular econometric computer packages such as *Microfit 4.0* and *PCGive 10.0.*

It is a simple matter to extend the algorithm of the previous section to produce a rolling regression. The additional task is to remove the data which was acquired at time $t-n$. The first step is to adjust the moment matrix to give $P_t^{*-1} = P_{t-1}^{-1} - x_{t-n}x_{t-n}'$. The matrix inversion formula of (A.3.ii) indicates that

$$\begin{aligned} P_t^* &= (P_{t-1}^{-1} - x_{t-n}x_{t-n}')^{-1} \\ &= P_{t-1} + P_{t-1}x_{t-n}(x_{t-n}'P_{t-1}x_{t-n} - 1)^{-1}x_{t-n}'P_{t-1}. \end{aligned} \tag{48}$$

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

$$\begin{aligned} b_t^* &= b_{t-1} - P_t^* x_{t-n}(y_{t-n} - x_{t-n}'b_{t-1}) \\ &= b_{t-1} - P_{t-1}x_{t-n}(x_{t-n}'P_{t-1}x_{t-n} - 1)^{-1}(y_{t-n} - x_{t-n}'b_{t-1}). \end{aligned} \tag{49}$$

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 (49). The second expression depends upon 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 the formula under (7) by replacing b_{t-1} with b_t^* and P_{t-1} with P_t^*.

The method of rolling regression is useful in initialising an ordinary recursive regression which lacks prior information on 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 can be replaced by sample information until the point $t = k$ is reached where there is only sample information in the data window. At that point, the rolling regression can be converted to an ordinary recursive regression; and the current values will be $b_k = X_k^{-1}Y_k$ and $P_k = (X_k'X_k)^{-1}$. In effect, 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.

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

Let $\lambda \in (0, 1]$ be the factor by which the data is discounted from one period to the next. Then, in place of the expression for P_t under (9), we should have

$$\begin{aligned} P_t &= (\lambda P_{t-1}^{-1} + x_t x_t')^{-1} \\ &= \frac{1}{\lambda}\Big\{P_{t-1} - P_{t-1}x_t(x_t'P_{t-1}x_t + \lambda)^{-1}x_t'P_{t-1}\Big\}. \end{aligned} \tag{50}$$

The formula for the parameter estimate would be

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + \lambda)^{-1}(y - x_t'b_{t-1}). \tag{51}$$

Discounted regression has yet to achieve widespread use in econometrics. It has been used extensively in the area of adaptive control, beginning with Åström, Borrison, Ljung and Wittenmark (1977). Its purpose, in this context, has been to prevent the recursive estimator from converging and to accommodate the drift in the parameters that characterise the system that is subject to control. A good example of an application is provided by Kiparissides and Shah (1983). Wellstead and Zarrop (1991) also give several practical examples.

Lozano (1983) has provided 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. For, in that case, the old information is forgotten while very little new information is added. This may make the control system over-sensitive to disturbances and susceptible to numerical and computational difficulties. The symptom of such difficulties is an explosive growth in the values within the dispersion matrix of the regression estimate.

The solution to the problem has been to devise systems of variable forgetting factors aimed at maintaining a constant information content within successive estimates. Analysis of such systems had been provided Zarrop (1983), Sanoff and Wellstead (1983) and Canetti and España (1989); and Fortescue, Kershenbaum and Ydstie (1981) have described an implementation. More sophisticated memory shaping systems are possible which will allow the information content to grow indefinitely, if there is no hint of parametric inconstancy, and which will discard information rapidly when there is clear evidence of change.

Apart from a belief in the parametric constancy of economic systems, there are several reasons that may be suggested for why econometricians have proved resistant to such devices as discounted regression. The first reason must be that, whereas occasional structural breaks can be accommodated easily, continuous structural change is liable to subvert the very objectives of structural econometric analysis. A second reason, which affects rolling regression as much as discounted regression, is that such devices are incapable of producing estimates that are statistically consistent. However, as we have indicated, this objection may be overcome by the use of sophisticated memory shaping.

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

6. The Kalman Filter

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

$$(52)\qquad y_t = H_t\beta_t + \eta_t, \qquad \textit{Observation Equation}$$

$$(53)\qquad \beta_t = \Phi_t\beta_{t-1} + \nu_t, \qquad \textit{Transition Equation}$$

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

$$(54)\qquad D(\eta_t) = \Omega_t \qquad \text{and} \qquad D(\nu_t) = \Psi_t.$$

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

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 time $t = 0$ 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:

$$b_{t|t-1} = \Phi_t b_{t-1}, \qquad \textit{State Prediction} \tag{55}$$

$$P_{t|t-1} = \Phi_t P_{t-1}\Phi_t' + \Psi_t, \qquad \textit{Prediction Dispersion} \tag{56}$$

$$e_t = y_t - H_t b_{t|t-1}, \qquad \textit{Prediction Error} \tag{57}$$

$$F_t = H_t P_{t|t-1} H_t' + \Omega_t, \qquad \textit{Error Dispersion} \tag{58}$$

$$K_t = P_{t|t-1} H_t' F_t^{-1}, \qquad \textit{Kalman Gain} \tag{59}$$

$$b_t = b_{t|t-1} + K_t e_t, \qquad \textit{State Estimate} \tag{60}$$

$$P_t = (I - K_t H_t) P_{t|t-1}. \qquad \textit{Estimate Dispersion} \tag{61}$$

It will also prove helpful to define

$$\Lambda_t = (I - K_t H_t)\Phi_t. \tag{62}$$

In comparison with the equations of the recursive regression algorithm listed under (10)–(14), there are two additions: equation (55) for the state prediction and equation (56) for its dispersion. These owe their existence to the presence of the transition equation (53); and they vanish when $\Phi = I$ and when $\nu_t = 0$ and $D(\nu_t) = \Psi_t = 0$, in which case $P_{t|t-1}$ becomes P_{t-1} in the remaining equations.

The equations of the Kalman filter may be derived using the results from the algebra of conditional expectations which are listed under (A.8).

Of the equations listed under (55)–(61), those under (57) and (59) are merely definitions.

To demonstrate equation (55), we use (A.8.iii) to show that

$$\begin{aligned} E(\beta_t|\mathcal{I}_{t-1}) &= E\big\{E(\beta_t|\beta_{t-1})|\mathcal{I}_{t-1}\big\} \\ &= E\big\{\Phi_t\beta_{t-1}|\mathcal{I}_{t-1}\big\} \\ &= \Phi_t b_{t-1}. \end{aligned} \tag{63}$$

We use (A.8.v) to demonstrate equation (56):

$$\begin{aligned} D(\beta_t|\mathcal{I}_{t-1}) &= D(\beta_t|\beta_{t-1}) + D\big\{E(\beta_t|\beta_{t-1})|\mathcal{I}_{t-1}\big\} \\ &= \Psi_t + D\big\{\Phi_t\beta_{t-1}|\mathcal{I}_{t-1}\big\} \\ &= \Psi_t + \Phi_t P_{t-1}\Phi_t'. \end{aligned} \tag{64}$$

To obtain equation (58), we substitute (52) into (57) 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, we have

$$\begin{aligned} D(e_t) &= D\big\{H_t(\beta_t - b_{t|t-1})\big\} + D(\eta_t) \\ &= H_t P_{t|t-1} H_t' + \Omega_t = D(y_t|\mathcal{I}_{t-1}). \end{aligned} \tag{65}$$

To demonstrate the updating equation (60), we begin by noting that

$$\begin{aligned} C(\beta_t, y_t|\mathcal{I}_{t-1}) &= E\big\{(\beta_t - b_{t|t-1})y_t'\big\} \\ &= E\big\{(\beta_t - b_{t|t-1})(H_t\beta_t + \eta_t)'\big\} \\ &= P_{t|t-1}H_t'. \end{aligned} \tag{66}$$

It follows from (A.8.i) that

$$\begin{aligned} E(\beta_t|\mathcal{I}_t) &= E(\beta_t|\mathcal{I}_{t-1}) + C(\beta_t, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\big\{y_t - E(y_t|\mathcal{I}_{t-1})\big\} \\ &= b_{t|t-1} + P_{t|t-1}H_t'F_t^{-1}e_t. \end{aligned} \tag{67}$$

The dispersion matrix under (61) for the updated estimate is obtained via equation (A.8.ii):

$$\begin{aligned} D(\beta_t|\mathcal{I}_t) &= D(\beta_t|\mathcal{I}_{t-1}) - C(\beta_t, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})C(y_t, \beta_t|\mathcal{I}_{t-1}) \\ &= P_{t|t-1} - P_{t|t-1}H_t'F_t^{-1}H_tP_{t|t-1}. \end{aligned} \tag{68}$$

It will be helpful for later analysis to express the current state vector in terms of the initial state vector and a sequence of state disturbances. Thus, by repeated back substitution in equation (53), we obtain

$$\beta_t = \sum_{j=1}^{t}\Phi_{t,j+1}\nu_j + \Phi_{t,1}\beta_0, \tag{69}$$

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 (52) gives another useful expression:

$$\begin{aligned} y_t &= H_t\Phi_{t,1}\beta_0 + H_t\sum_{j=1}^{t}\Phi_{t,j+1}\nu_j + \eta_t \\ &= X_t\beta_0 + \varepsilon_t. \end{aligned} \tag{70}$$

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

$$y = X\beta_0 + \varepsilon, \quad \text{where} \quad E(\varepsilon) = 0 \quad \text{and} \quad D(\varepsilon) = \Sigma. \tag{71}$$

The remaining task of this section is to establish 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 (55) and (57) into (60) to give

$$\begin{aligned} b_t &= \Phi_tb_{t-1} + K_t(y_t - H_t\Phi_tb_{t-1}) \\ &= \Lambda_tb_{t-1} + K_ty_t, \end{aligned} \tag{72}$$

where we have used $\Lambda_t = (I - K_tH_t)\Phi_t$ from (62). Repeated back-substitution gives

$$b_t = \sum_{j=1}^{t} \Lambda_{t,j+1}K_jy_j + \Lambda_{t,1}b_0, \tag{73}$$

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

$$\begin{aligned} e_t &= y_t - H_t\Phi_tb_{t-1} \\ &= y_t - H_t\Phi_t \sum_{j=1}^{t-1} \Lambda_{t-1,j+1}K_jy_j - H_t\Phi_t\Lambda_{t-1,1}b_0, \end{aligned} \tag{74}$$

which is a straightforward generalisation of equation (41). On defining the vector $e = [e'_1, \ldots, e'_T]'$, the T equations can be written as

$$e = Ly - Wb_0 = L(y - Xb_0), \quad \text{with} \quad E(e) = 0 \quad \text{and} \quad D(e) = F. \tag{75}$$

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 (74) 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 (55) and (60), we get

$$b_{t-1} = \sum_{j=1}^{t-1} \Phi_{t-1,j+1}K_je_j + \Phi_{t-1,1}b_0. \tag{76}$$

Substituting $b_{t|t-1} = \Phi_tb_{t-1}$ into equation (57) gives

$$\begin{aligned} y_t &= e_t + H_tb_{t|t-1} \\ &= e_t + H_t \sum_{j=1}^{t-1} \Phi_{t,j+1}K_je_j + H_t\Phi_{t,1}b_0. \end{aligned} \tag{77}$$

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.

7. 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 which must be approached with care. The present section lays the necessary groundwork.

We have 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 (71), where the disturbances are normally distributed with $E(\varepsilon) = 0$ and $D(\varepsilon) = \Sigma$. There is $E(y) = XE(\beta_0) + E(\varepsilon)$ and $D(y) = XD(\beta_0)X' + D(\varepsilon)$. Therefore,

$$E(y) = Xb_0, \tag{78}$$

$$D(y) = XP_0X' + \Sigma, \tag{79}$$

$$E(\beta_0) = b_0, \tag{80}$$

$$D(\beta_0) = P_0, \tag{81}$$

$$C(y, \beta_0) = XP_0. \tag{82}$$

The joint density function of y and β_0 is

$$N(y, \beta_0) = (2\pi)^{-(T+k)/2}|D(y, \beta_0)|^{-1/2} \exp\{-S(y, \beta_0)/2\}, \tag{83}$$

of which, according to (A.6), the quadratic of the exponent can be written variously as

(84)

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

In the final expression, the identity

$$P_0 - P_0X'(XP_0X' + \Sigma)^{-1}XP_0 = (X'\Sigma^{-1}X + P_0^{-1})^{-1}, \tag{85}$$

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 (84), 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) from the appendix. It is given by

$$\begin{aligned} E(\beta_0|y) &= b_0 + P_0X'(XP_0X' + \Sigma)^{-1}(y - Xb_0) \\ &= b_0 + (X'\Sigma^{-1}X + P_0^{-1})^{-1}X'\Sigma^{-1}(y - Xb_0) \\ &= (X'\Sigma^{-1}X + P_0^{-1})^{-1}(X'\Sigma^{-1}y + P_0^{-1}b_0) = b_*, \end{aligned} \tag{86}$$

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

$$P_0X'(XP_0X'+\Sigma)^{-1}=(X'\Sigma^{-1}X+P_0^{-1})^{-1}X'\Sigma^{-1}. \tag{87}$$

(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 (84) can be written in a summary notation as

$$\begin{aligned}S(y,\beta_0)&=S(y|\beta_0)+S(\beta_0)\\&=S(\beta_0|y)+S(y),\end{aligned} \tag{88}$$

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:

$$S(y|\beta_0)=(y-X\beta_0)'\Sigma^{-1}(y-X\beta_0), \tag{89}$$

$$S(\beta_0)=(\beta_0-b_0)'P_0^{-1}(\beta_0-b_0), \tag{90}$$

$$S(\beta_0|y)=(\beta_0-b_*)'(X'\Sigma^{-1}X+P_0^{-1})(\beta_0-b_*), \tag{91}$$

$$\begin{aligned}S(y)&=(y-Xb_0)'(XP_0X'+\Sigma)^{-1}(y-Xb_0)\\&=(y-Xb_0)'\{\Sigma^{-1}-\Sigma^{-1}X(X'\Sigma^{-1}X+P_0^{-1})^{-1}X'\Sigma^{-1}\}(y-Xb_0).\end{aligned} \tag{92}$$

The second expression for $S(y)$ on the RHS of (92) 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

$$|P_0|=|XP_0X'+\Sigma||X'\Sigma^{-1}X+P_0^{-1}|^{-1}. \tag{93}$$

The various ways of estimating β_0 can be considered in the light of the foregoing algebraic results. First to be considered is the estimator obtained by maximising, in respect of β_0, the likelihood function corresponding to the conditional density function $N(y|\beta_0)$. When this approach is taken, the tendency is to regard β_0 as a parametric constant, as opposed to the realised value of a random variable, in which case, the *conditional* likelihood function becomes the *unconditional* function. The result, which is obtained, in any case, by minimising the quadratic function $S(y|\beta_0)$ of (89), will be described as the unconditional estimator:

$$b_{0|T}=(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y. \tag{94}$$

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

$$N^c(y)=(2\pi)^{-T/2}|\Sigma|^{-1/2}\exp\{-S^c(y)/2\}, \tag{95}$$

wherein

$$S^c(y)=y'\{\Sigma^{-1}-\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\}y. \tag{96}$$

A purpose of defining the concentrated function is to provide 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.

The next estimator of the initial state vector is its conditional expectation $b_* = E(\beta_0|y)$, specified in alternative forms by equation (86). This estimator is also derivable 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 $N(y, \beta_0)$. Letting $P_0 \rightarrow \infty$ in (86), which is tantamount to negating the priori information on β_0, results in the unconditional estimator $b_{0|T}$ of (94), which is 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

$$N(y) = (2\pi\sigma)^{-T/2}|XP_0X' + \Sigma|^{-1/2}\exp\{-S(y)/2\}, \tag{97}$$

wherein the quadratic exponent $S(y)$ is given by (92). Differentiating $S(y)$ with respect to b_0 and setting the result to zero gives a first-order condition from which is obtained the maximum-likelihood estimator

$$\begin{aligned}\hat{b}_0 &= \{X'(XP_0X' + \Sigma)^{-1}X\}^{-1}X'(XP_0X' + \Sigma)^{-1}y \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y = b_{0|T}.\end{aligned} \tag{98}$$

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 $(y; X\beta, \Omega_2)$ will be identical if and only if the columns of the matrices $\Omega_1^{-1}X$ and $\Omega_1^{-1}X$ span the same space—see Pollock (1979, p. 86), for example.) However, the equality can be demonstrated directly by reference to (87), 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 (98), 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 of which the quadratic exponent is the function $S^c(y)$ of (96). The likelihood can be maximised further by setting $P_0 = 0$. The result is, once more, the function $N^c(y)$ of (95). 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 $P_0 \rightarrow \infty$ within the marginal distribution $N(y)$ of (97) to create what de Jong (1988a, 1991) and Ansley and Kohn (1985a, 1986, 1990) have described as a diffuse distribution. The effect within the exponent is that $S(y) \rightarrow S^c(y)$. The effect within the determinantal term is problematic, since XP_0X' is unbounded. However, in view of (93), 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 to define the diffuse likelihood function by

$$N^d(y) = |X'\Sigma^{-1}X|^{-1/2}(2\pi)^{-T/2}\exp\{-S^c(y)/2\}. \tag{99}$$

The quadratic exponent $S^c(y)$ of the diffuse likelihood, which is the essential part, is identical to the one which arises from concentrating the marginal likelihood function $N(y)$ of (97) 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, when negating the prior information, by letting $P_0 \to \infty$, it is best to do so in the context of the joint distribution factorised as $N(y, \beta_0) = N(y|\beta_0)N(\beta_0)$. For this allows the difficulties of the limiting process to be confined to the factor $N(\beta_0)$.

Example. There are several alternative ways of deriving an expression for the quadratic component of the marginal distribution $N(y)$ which lead to expressions which are so markedly different that one must struggle to 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 $N(y, \beta_0) = N(\beta_0|y)N(y)$ will deliver $S(y)$, since the term $S(\beta_0|y)$ is thereby eliminated. The 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

$$S(y) = (b_* - b_0)'P_0^{-1}(b_* - b_0) + (y - Xb_*)'\Sigma^{-1}(y - Xb_*). \tag{100}$$

The expression has been exploited by Goméz and Maravall (1994a). This procedure for finding $S(y)$ has also been followed by Box and Jenkins (1976) in pursuit of 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

$$\begin{aligned} S(y) = {} & (y - X\beta_0)'\Sigma^{-1}(y - X\beta_0) + (\beta_0 - b_0)'P_0^{-1}(\beta_0 - b_0) \\ & - (\beta_0 - b_*)'(X'\Sigma^{-1}X + P_0^{-1})(\beta_0 - b_*). \end{aligned} \tag{101}$$

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 of the expressions of (100) and (101) 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.

8. Transformations and the Problem of Initialisation

In the econometric literature, there has been a tendency to adopt the transformations approach in dealing the initialisation problem that occurs when the Kalman filter is applied to a nonstationary process. This, undoubtably, reflects the influence of Ansley and Kohn (1985a). The purpose of the transformation is to eliminate the dependence of the likelihood upon the unknown initial values. It has also been customary to illustrate solutions to the problem by reference to the likelihood function of an autoregressive integrated moving-average (ARIMA) model.

Confusion over the transformations approach can arise from the fact that it may be used as a theoretical device when there is no intention of applying it in practice. Indeed, in devising their modified Kalman filter, Ansley and Kohn (1985a) sought to avoid transformations of the data which would obstruct their handling of the problem of missing observations.

To illustrate the approach of Ansley and Kohn, let us consider the orthonormal matrix $C = [C_1, C_2]$, defined in section 4 in connection with the LUS residuals. The columns of C_1 span the same space as the columns of X, whereas $C_2'X = 0$. Therefore, transforming $y = X\beta_0 + \varepsilon$ by C' gives

$$(102)\qquad \begin{bmatrix} C_1'y \\ C_2'y \end{bmatrix} = \begin{bmatrix} C_1'X\beta_0 \\ 0 \end{bmatrix} + \begin{bmatrix} C_1'\varepsilon \\ C_2'\varepsilon \end{bmatrix},$$

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

$$(103)\qquad S^c(y) = y'C_2(C_2'\Sigma C_2)^{-1}C_2'y = y'\{\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\}y,$$

of which the RHS is identical to the expression under (96), which represents the quadratic exponent of both the concentrated likelihood function $N^c(y)$ and the diffuse likelihood function $N^d(y)$. The second equality of (103) follows from the fact that, if $\text{Rank}[W, X] = T$ and if $W'\Sigma^{-1}X = 0$, then

$$(104)\qquad W(W'\Sigma^{-1}W)^{-1}W'\Sigma^{-1} = I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}.$$

The equality is obtained by premultiplying both sides of (104) by Σ^{-1} and then setting $W = \Sigma C_2$. Observe that, when $\Sigma = I$, equation (103) specialises to equation (46), which represents the sum of squares of the LUS residuals of the ordinary regression model.

An alternative transformation has been proposed by Bell and Hillmer (1991). They set $X = [X_1', X_2']'$ and $y = [y_1', y_2']'$, where X_1 and y_1 comprise the first k observations, where k is the dimension of β_0. Then, they apply the following transformation:

$$(105)\qquad \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} X_1^{-1} & 0 \\ -X_2X_1^{-1} & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ 0 \end{bmatrix} + \begin{bmatrix} X_1^{-1}\varepsilon_1 \\ -X_2X_1^{-1}\varepsilon_1 + \varepsilon_2 \end{bmatrix}.$$

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

$$(106)\qquad S^c(y) = z_2'D^{-1}(z_2)z_2 = (y_2 - X_2b_{0|k})'D^{-1}(z_2)(y_2 - X_2b_{0|k})$$

is an alternative representation of the concentrated sum of squares. This expression is analogous to equation (44), which relates to ordinary recursive regression. 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 (106) identical to (44).

In order to apply the transformation approach to an ARIMA process, one must begin by demonstrating the dependence of such a process on its initial conditions. The ARIMA process may be represented by

$$(107)\qquad \alpha(L)\delta(L)y(t) = \mu(L)\varepsilon(t),$$

where $\alpha(z)$ and $\mu(z)$ are, respectively, the autoregressive and the moving-average polynomials, which have their roots outside the unit circle, and where $\delta(z)$ is a polynomial of degree d, which has roots of unit modulus. The equation can also be written as $\delta(L)y(t) = \{\mu(L)/\alpha(L)\}\varepsilon(t) = \zeta(t)$, where $\zeta(t)$ is a stationary ARMA process.

Let $I = [e_1, e_2, \ldots, e_T]$ be the identity matrix of order T, from which $K = [e_2, \ldots, e_T, e_1]$ and $L = [e_2, \ldots, e_T, 0]$ are derived. Replacing the argument z in the polynomial $\delta(z)$ by K gives a circulant matrix $\delta(K) = \Gamma = \Delta + \nabla$, which is the sum of a lower-triangular matrix $\Delta = \delta(L)$ and a complementary upper-triangular matrix $\nabla = \delta(K - L)$. We shall let ∇_* denote the matrix consisting of the last d columns of ∇, which is where all of its non-zero elements are to be found.

To form the matrix representation, let $y_* = [y_{1-d}, \ldots, y_0]'$ be a vector of d presample elements of $y(t)$ and let $\zeta = [\zeta_1, \ldots, \zeta_T]'$ contain the elements of the ARMA process within the sample period. Then the observations on the ARIMA process are the elements of the vector $y = [y_1, \ldots, y_T]'$, which is to be found within the following equations:

$$(108)\quad \text{(i)}\quad \begin{bmatrix} I & 0 \\ \nabla_* & \Delta \end{bmatrix} \begin{bmatrix} y_* \\ y \end{bmatrix} = \begin{bmatrix} y_* \\ \zeta \end{bmatrix} \quad \text{(ii)}\quad \begin{bmatrix} y_* \\ y \end{bmatrix} = \begin{bmatrix} I & 0 \\ -\Delta^{-1}\nabla_* & \Delta^{-1} \end{bmatrix} \begin{bmatrix} y_* \\ \zeta \end{bmatrix}.$$

Equation (ii), which is obtained from equation (i) by inverting the matrix, shows that the vector $y = \Delta^{-1}\zeta - \Delta^{-1}\nabla_* y_*$ of the observations of the ARIMA process depends upon the initial conditions of y_* and on the vector ζ, which is generated by a stationary ARMA process. Bell (1984), for example, has used of this result in discussing the filtering of nonstationary sequences.

There are now two ways of tackling the initial-value problem. The approach of Ansley and Kohn (1985a) is to work with the marginal distribution of y. The vector y_* of initial conditions is mapped into the initial state vector β_0. The result is a diffuse random vector compounded from diffuse and non-diffuse elements. Ansley and Kohn have devised a modified Kalman filter in which both the diffuse and the non-diffuse information is used in estimating the state vectors from $t = 1$ to $t = d - 1$. As each new observation is assimilated, an element of diffuse information is replaced until, at time $t = d = k$, the estimate becomes the product of sample information alone.

The modified Kalman filter, which performs its iterations from the start, represents a sophisticated means of boot-strapping the filtering process. Ansley and Kohn justify their approach by showing that it produces the same results as a transformation approach that eliminates the effect of the starting values. Their means of demonstrating this proposition is to show that, when $\rho \to \infty$ within $P_0 = \rho I$, the marginal likelihood function $N(y)$ of (97) becomes the diffuse likelihood function $N^d(y)$ of (99) which is, in essence, the likelihood function of the transformed vector $C_2'y$. Refinements to the modified Kalman filter has been published by Ansley and Kohn (1990), and detailed descriptions of its use in estimating nonstationary ARIMA models have been given by Ansley and Kohn (1985b) and Kohn and Ansley (1986).

The alternative way of handling the initial-value problem of the ARIMA model, which sacrifices the first d iterations of the filter, is to work with the conditional likelihood function. Consider

$$\nabla_* y_* + \Delta y = \begin{bmatrix} \nabla_{1*} \\ 0 \end{bmatrix} y_* + \begin{bmatrix} \Delta_{11} & 0 \\ \Delta_{21} & \Delta_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}, \tag{109}$$

where y_*, y_1 and ζ_1 are all of order d. Then, by analogy with equation (107.i), there is

$$\begin{bmatrix} I & 0 \\ \Delta_{21} & \Delta_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ \zeta_2 \end{bmatrix}. \tag{110}$$

The matrix on the LHS, which has units of the diagonal and zeros above, is of full rank and it has a determinant of unit value. Therefore, it follows that $N(y_1, y_2) = N(y_1, \zeta_2)$. But $N(y_1, y_2) = N(y_1)N(y_2|y_1)$ and $N(y_1, \zeta_2) = N(y_1)N(\zeta_2)$, so there is $N(y_2|y_1) = N(\zeta_2)$. Moreover, $\zeta_2 = \Delta_{21}y_1 + \Delta_{22}y_2$. It follows that, once the starting values in y_1 have been acquired, the elements of ζ_2 can be calculated, at which point the conditional likelihood function $N(y_2|y_1)$ of the ARIMA model becomes synonymous with the likelihood function $N(\zeta_2)$ of the stationary ARMA model.

These observations are due to Maravall and Goméz (1994b). The same prescriptions are to be found in the paper of Bell and Hillmer (1991), where they follow directly from the transformation represented by equation (105). The paper also treats the unobserved ARIMA components model, in which respect it may be compared with the paper of Kohn and Ansley (1987), which employs their modified Kalman filter.

9. Calculating the Estimate of the Initial State

There are various ways in which, in practice, the values of $\mathcal{I}_0 = \{b_0, P_0\}$ might be obtained, which are used in starting up the Kalman filter. Often, the assumption that the state vectors are generated by a stationary process can be used in finding analytic expressions for b_0 and P_0. Under the assumption of stationarity, the matrices H_t, Φ_t, Ω_t and Ψ_t become constant, and they loose their temporal subscripts.

For stationarity, the eigenvalues of the transformation matrix Φ must lie within the unit circle, which implies that $\lim(n \to \infty)\Phi^n = 0$. In that case, the unconditional moments $E(\beta_0) = b_0 = 0$ and $D(\beta_0) = P_0 = \Phi P_0 \Phi' + \Psi$, which come from equation (53), 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 by pursuing a convergent iterative process, of which the ith step is described by $P_i = \Phi P_{i-1}\Phi' + \Psi$.

In the case where the state space equations (52) and (53) 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), for example. There are also ways of formulating the state-space representation of the ARMA model that facilitate the direct derivation of the matrix P_0. Such methods have been described by Mittnik (1987a, 1987b) and by Diebold (1986a, 1986b).

When the state vectors are generated by a non-stationary process, the initial vector β_0 is liable to have an unknown distribution. Then an estimate of b_0 can be found by maximising a likelihood function which is commonly obtained from the marginal distribution of $N(y)$, of which the quadratic form can be written as

$$\begin{aligned} S(y) &= (y - Xb_0)'(XP_0X' + \Sigma)^{-1}(y - Xb_0) \\ &= (y - Xb_0)'L'F^{-1}L(y - Xb_0) = e'F^{-1}e, \end{aligned} \tag{111}$$

where F is a block-diagonal matrix with F_t as the tth diagonal block. Here, the first expression on the RHS is from (92), whereas the second expression, which reflects the identities of (75), is the form proposed originally by Schweppe (1965).

It has been show, in the preceding section, that the value that minimises $S(y)$ is the estimator $b_{0|T} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ of (94), 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 (89).

(Setting $P_0 = 0$, in this context does not carry the literal interpretation that β_0 is now know 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 (94) is not directly amenable to computation. To derive an operational form, consider writing equation (74) as

$$\begin{aligned} e_t &= \Big\{y_t - H_t\Phi_t\sum_{j=1}^{t-1}\Lambda_{t-1,j+1}K_jy_j\Big\} - H_t\Phi_t\Lambda_{t-1,1}b_0 \\ &= e_t^* - W_tb_0, \end{aligned} \tag{112}$$

where e_t^* and W_tb_0 are the tth subvectors, respectively, of Ly and $Wb_0 = LXb_0$, which are to be found in equation (75). Substituting in $S(y) = \sum_{t=1}^{T} e_t'F^{-1}e_t$ gives

$$S(y) = \sum_{t=1}^{T}(e_t^* - W_tb_0)'F_t^{-1}(e_t^* - W_tb_0). \tag{113}$$

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

$$b_{0|T} = \Big(\sum_{t=1}^{T}W_t'F_t^{-1}W_t\Big)^{-1}\sum_{t=1}^{T}W_t'F_t^{-1}e_t^* = M_T^{-1}m_T. \tag{114}$$

The elements of this expression can be accumulated via the recursions

$$\begin{aligned} m_t &= m_{t-1} + \Lambda_{t-1,1}'\Phi_t'H_t'F_t^{-1}e_t^*, \\ M_t &= M_{t-1} + \Lambda_{t-1,1}'\Phi_t'H_t'F_t^{-1}H_t\Phi_t\Lambda_{t-1,1}, \end{aligned} \tag{115}$$

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

$$\Lambda_{t,1} = (\Phi_t - K_tH_t\Phi_t)\Lambda_{t-1,1}, \tag{116}$$

which starts with $\Lambda_{1,1} = \Lambda_1$. Notice, however, in reference to (112), 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 by replacing the observations y_t by zeros, the sequence $\{W_t\}$ is generated.

The objective of estimating the initial conditions can therefore be accomplished by extending two of the equations of the Kalman filter and by adding an extra equation. Thus,

$$E_t = Y_t - H_t\Phi_t B_{t-1}, \qquad \textit{Extended Prediction Error} \tag{117}$$

$$B_t = \Phi_t B_{t-1} + K_t E_t, \qquad \textit{Extended State Estimate} \tag{118}$$

$$Q_t = Q_{t-1} + E_t' F_t^{-1} E_t. \qquad \textit{Cross - Product Accumulation} \tag{119}$$

Here, equations (117) and (118) are extensions of (57) and (60) 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, predicated upon 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 $Q_0 = 0$. The matrix Q_t is as follows:

$$Q_t = \begin{bmatrix} S_t & m_t \\ m_t' & M_t \end{bmatrix}. \tag{120}$$

This contains the quantities defined in (115) together with the sum of squares of the prediction errors scaled by their variance.

The algorithm that we have described is attributable to Rosenberg (1973). It has been expounded by Harvey (1989), amongst others, and de Jong (1988a, 1988b, 1989, 1991a, 1991b) has used it in a succession of papers. See also de Jong and Chu-Chun-Lin, (1994, 2002).

There are various strategies that can be followed in assimilating the estimated starting values to the state estimates. 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, thereafter, to adjust them using the estimate $b_{0|T}$ of (114). It follows from (73) that the adjusted estimate of β_t is $b_t = b_t^* + \Lambda_{t,1} b_{0|T}$.

An alternative procedure is to collapse the extended filter by absorbing a tentative estimate of the starting value into the state estimate and proceeding with the standard Kalman filter. This suggestion has been made by de Jong (1991a, 1991b), and it is in accordance with the prescriptions of Bell and Hillmer (1991).

The earliest opportunity of collapsing the filter arises when the $k \times k$ matrix M_t first achieves a rank of k, which, in the case of univariate observations, it is liable to do when $t = k$. 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 will have values that would be given by

$$\begin{aligned} e_t &= e_t^* - W_t M_t^{-1} m_t, \\ b_t &= b_t^* + \Lambda_{t,1} M_t^{-1} m_t, \end{aligned} \tag{121}$$

if one were to calculate the quantities on the RHS. Thus, the standard Kalman filter will implicitly enhance 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 will be

$$\begin{aligned} D(b_t) &= D(b_t^*) + \Lambda_{t,1} D(b_{0|T})\Lambda_{t,1}' \\ &= P_t^* + \Lambda_{t,1} M_t^{-1}\Lambda_{t,1}' = P_t; \end{aligned} \tag{122}$$

and this will also be generated directly the standard (collapsed) filter—see de Jong and Chu-Chun-Lin (1994).

A problem which arises from collapsing filter is how to find estimates for the state vectors $\beta_1, \ldots, \beta_{k-1}$ which occur prior to the collapse when the first estimate of the starting value is formed. One evident solution, which is outlined by de Jong and Chu-Chun-Lin (2002), is to use the estimate $b_{0|k} = M_k^{-1}m_k$ to adjust the pre-collapse values in the manner that $b_{0|T} = M_T^{-1}m_T$ is used in Rosenberg's procedure. Although the resulting state estimates will be based on a tenuous estimate of the starting value, they can be improved, nevertheless, in a subsequent smoothing operation.

The smoothed estimates of the state vectors will not be affected by the matter of whether $b_{0|k}$ or $b_{0|T}$ has been incorporated in preliminary estimates obtained from filtering. Smoothing adds any information that is missing from the estimates, but it has no effect if the information has been incorporated already.

The essential features of a quite different method of initialising the filter that is due to Ansley and Kohn (1985a) have already been presented at the end of section 3 in the context of an ordinary recursive regression. The method depends upon setting $P_t = P_t^* + \rho P_t^\circ$, where P_t° relates to the diffuse component of the prior information and where $\rho \to \infty$. For the case where $P_t^\circ > 0$ and $f_t^\circ > 0$, the algorithm has been 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 in order to apply the method in the present context. First, there is $P_{t|t-1} = P_{t|t-1}^* + \rho P_{t|t-1}^\circ$, where

$$P_{t|t-1}^\circ = \Phi_t P_{t-1}^\circ \Phi_t' \qquad \text{and} \qquad P_{t|t-1}^* = \Phi_t P_{t|t-1}^* \Phi_t' + \Psi_t. \tag{123}$$

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

$$F_t^\circ = H_t P_{t|t-1}^\circ H_t' \qquad \text{and} \qquad F_t^* = H_t P_{t|t-1}^* H_t' + \Omega_t. \tag{124}$$

Usually, the assumption can be made that, if it is not zero-valued, then F_t° is non singular—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:

$$b_t = b_{t|t-1} + P_{t|t-1}^\circ H_t F_t^{\circ -1}(y_t - H_t x_{t|t-1}), \tag{125}$$

$$P_t^\circ = P_{t|t-1}^\circ - P_{t|t-1}^\circ H_t F_t^{\circ -1} H_t' P_{t|t-1}^\circ, \tag{126}$$

$$\begin{aligned} P_t^* = P_{t|t-1}^* &+ P_{t|t-1}^\circ H_t F_t^{\circ -1} F_t^* F_t^{\circ\prime -1} H_t' P_{t|t-1}^\circ, \\ &- P_{t|t-1}^\circ H_t F_t^{\circ -1} H_t' P_{t|t-1}^* - P_{t|t-1}^* H_t F_t^{\circ -1} H_t' P_{t|t-1}^\circ. \end{aligned} \tag{127}$$

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

$$b_t = b_{t|t-1} + P^*_{t|t-1} H_t F_t^{*-1}(y_t - H_t x_{t|t-1}), \tag{128}$$

$$P_t^\circ = P^\circ_{t|t-1}, \tag{129}$$

$$P_t^* = P^*_{t|t-1} - P^*_{t|t-1} H_t F_t^{*-1} H_t' P^*_{t|t-1}. \tag{130}$$

These are just the equations of the standard Kalman filter.

The original derivation by Ansley and Kohn (1985a) was somewhat laborious, and the subsequent abbreviated derivation by Kohn and Ansley (1986) is more accessible. A modified version of the algorithm, for which superior numerical accuracy is claimed, has been provided Ansley and Kohn (1990). Other derivations have been provided by Snyder (1988), who has considered a square-root version of the Kalman filter, and by Koopman (1997) who has treated the most general case where $F_t^\circ > 0$ is not necessarily a nonsingular matrix.

One virtue of the foregoing approach to initialising the filter is that it provides a complete sequence of state estimates and of their corresponding dispersion matrices for $t = 1, \ldots, T$ that is amenable to standard versions of the smoothing algorithms—see Koopman (1997).

10. The Smoothing Algorithms

The Kalman filter, which is commonly used as a real-time or on-line algorithm, creates estimates of the state vectors using current and past information. Often, there is scope for the enhancement of these estimates using information that has transpired subsequently.

In the digital processing of speech, prior to its transmission via the telephone, it is acceptable to impose a small delay for the purpose of gathering extra information. A fixed-lag smoothing algorithm can then be used to enhance the digital signal. In econometrics, where there is no immediate real-time constraint, it is possible to use all of the subsequent information within a given sample to enhance the state estimates. For this purpose, the so-called fixed-interval smoothing algorithms are appropriate.

Smoothing algorithms were quickly provided following the original publication of Kalman (1960). A notable contributor was Rauch (1963); and the early work was surveyed by Meditch (1973). Whereas the fixed-lag smoothing algorithms have featured prominently in the engineering literature, the fixed-interval algorithms have received less attention; and econometricians have found scope for developing them. Notable contributions have been by Ansley and Kohn (1982), Kohn and Ansley (1989), de Jong (1988b, 1989) and Koopman (1993). All classes of smoothing algorithms have been surveyed and compared by Merkus, Pollock and De Vos (1993). In this section, we shall concentrate exclusively on the fixed-interval algorithms. The essential task will be to find computable expressions for the covariances of the prediction errors and the state vectors.

Given that the prediction errors are mutually independent, it follows from

(A.8.i) that

$$E(\beta_t|\mathcal{I}_T) = E(\beta_t|\mathcal{I}_t) + \sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)e_j. \tag{131}$$

This represents the means by which the estimate $b_t = E(\beta_t|\mathcal{I}_t)$ is updated using the information $\{e_{t+1}, \ldots, e_T\}$ that has arisen subsequent to time t in order to produce the definitive estimate $b_{t|T} = E(\beta_t|\mathcal{I}_T)$. It also follows from (A.8.ii) that the dispersion matrix of the estimate is

$$D(\beta_t|\mathcal{I}_T) = E(\beta_t|\mathcal{I}_t) - \sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)C(e_j, \beta_t). \tag{132}$$

The task in realising these equations is to devise a recursive scheme which will produce the sequence of updated estimates in an appropriate order and in a way which minimises the necessary calculations each stage.

Consider

$$e_k = H_k\Phi_k(\beta_{k-1} - b_{k-1}) + H_k\nu_k + \eta_k, \tag{133}$$

which comes from substituting the transition equation (53) into the observation equation (52) to give $y_k = H_k(\Phi_k\beta_{k-1} + \nu_k) + \eta_k$ and thereafter subtracting $H_kb_{k|k-1} = H_k\Phi_kb_{k-1}$. Within this expression, there is

$$\beta_{k-1} - b_{k-1} = \Lambda_{k-1}(\beta_{k-2} - b_{k-2}) + (I - K_{k-1}H_{k-1})\nu_{k-1} - K_{k-1}\eta_{k-1}. \tag{134}$$

This equation is obtained by subtracting $b_{k-1} = \Phi_{k-1}b_{k-2} + K_{k-1}e_{k-1}$ from the transition equation and thereafter by substituting the expression for e_{k-1} from (133) into the result. The equation is amenable to a recursion. Running the recursion from $k-1$ down to t gives

$$\beta_{k-1} - b_{k-1} = \Lambda_{k-1,t+1}(\beta_t - b_t) + \sum_{j=t+1}^{k-1} \Lambda_{k-1,j+1}\{(I - K_jH_j)\nu_j - K_j\eta_j\}. \tag{135}$$

The terms under the summation comprise stochastic elements that are subsequent to t and which are therefore independent of the prediction error e_t. After drafting (135) into (133), It follows that, when $k > t$, there is

$$\begin{aligned} C(\beta_t, e_k) &= E\{\beta_t(\beta_t - b_t)\Lambda'_{k-1,t+1}\Phi'_kH'_k\} \\ &= P_t\Lambda'_{k-1,t+1}\Phi'_kH'_k. \end{aligned} \tag{136}$$

Now consider

$$C(\beta_{t+1}, e_k) = P_{t+1}\Lambda'_{k-1,t+2}\Phi_kH_k. \tag{137}$$

The comparison of (136) and (137) shows that

$$\begin{aligned} C(\beta_t, e_k) &= P_t\Lambda'_{t+1}P^{-1}_{t+1}C(\beta_{t+1}, e_k) \\ &= P_t\Phi'_{t+1}P^{-1}_{t+1|t}C(\beta_{t+1}, e_k). \end{aligned} \tag{138}$$

Here, the identity $P^{-1}_{t+1}\Lambda_{t+1} = P^{-1}_{t+1|t}\Phi_{t+1}$, which gives the second equality, comes via (61) and (62), which indicate that $P_{t+1} = \Lambda_{t+1}\Phi^{-1}_{t+1}P_{t+1|t}$.

Equation (138) provides the recursion with which to implement the formulae of (131) and (132). The classical fixed-interval smoother is derived from the equation

$$E(\beta_t|\mathcal{I}_T) = E(\beta_t|\mathcal{I}_t) + P_t\Phi'_{t+1}P^{-1}_{t+1|t}\sum_{j=t+1}^{T} C(\beta_{t+1}, e_j)D^{-1}(e_j)e_j, \tag{139}$$

which is obtained by substituting the identity of (138) into equation (131). But

$$E(\beta_{t+1}|\mathcal{I}_T) = E(\beta_{t+1}|\mathcal{I}_t) + \sum_{j=t+1}^{T} C(\beta_{t+1}, e_j)D^{-1}(e_j)e_j, \tag{140}$$

so it follows that equation (139) can be rewritten in turn as

$$b_{t|T} = b_t + P_t\Phi'_{t+1}P^{-1}_{t+1|t}\{b_{t+1|T} - b_{t+1|t}\}, \tag{141}$$

where the notations $b_{t+1|T} = E(\beta_{t+1}|\mathcal{I}_T)$ 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 followed in deriving the dispersion matrix of the smoothed estimate. Corresponding to (139), there is

$$D(\beta_{t+1}|\mathcal{I}_T) = D(\beta_{t+1}|\mathcal{I}_t) - \sum_{j=t+1}^{T} C(\beta_{t+1}, e_j)D^{-1}(e_j)C(e_j, \beta_{t+1})e_j. \tag{142}$$

Therefore equation (132) can be written as

$$P_{t|T} = P_{t|t} + P_t\Phi'_{t+1}P^{-1}_{t+1|t}\{P_{t+1|T} - P_{t+1|t}\}P^{-1}_{t+1|t}\Phi_{t+1}P_t. \tag{143}$$

The classical formulae presuppose a process of forward filtering which generates the sequence $b_t; t = 1, \ldots, T$ of state estimates. The smoothing is realised by running backward through the sequence in a manner which entails a first-order feedback in respect of the smoothed estimates. The algorithm is due to Rauch (1963) and a derivation of it can be found in the text of Anderson and Moore (1979) and in many other sources.

In circumstances where the factor $P_t\Phi'_{t+1}P^{-1}_{t+1|t}$ can be represented by a constant matrix, the classical algorithm is efficient and easy to implement. This would be the case if there were a constant transition matrix Φ and if the filter gain K_t had converged to a constant. In all other circumstances, where it is required recompute the factor at each iteration of the index t, the algorithm is liable to cost time and to invite numerical inaccuracies. The problem, which lies with the inversion 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 equation (131), which, using (136), can be written as

$$\begin{aligned}&\sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)e_j\\ &\qquad = P_t \sum_{j=t+1}^{T} \Lambda'_{j-1,t+1}\Phi'_j H'_j F_j^{-1} e_j = P_t q_{t+1}.\end{aligned} \tag{144}$$

Also, within (132), there is

$$\begin{aligned}&\sum_{j=t+1}^{T} C(\beta_t, e_j)D^{-1}(e_j)C(\beta_t, e_j)\\ &\qquad = P_t\Big\{ \sum_{j=t+1}^{T} \Lambda'_{j-1,t+1}\Phi'_j H'_j F_j^{-1} H_j \Phi_j \Lambda_{j-1,t+1}\Big\}P_t = P_t Q_{t+1} P_t.\end{aligned} \tag{145}$$

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

$$\begin{aligned}q_t &= \Phi'_t H'_t F_t^{-1} e_t + \Lambda'_{t+1} q_{t+1},\\ Q_t &= \Phi'_t H'_t F_t^{-1} H'_t \Phi'_t + \Lambda'_{t+1} Q_{t+1} \Lambda'_{t+1},\end{aligned} \tag{146}$$

and which are initiated with $q_T = \Phi'_T H'_T F_T^{-1} e_T$ and $Q_T = \Phi'_T H'_T F_T^{-1} H_T \Phi_T$. Notice that these are the counterparts of the recursions of (114) which run forwards in time. The recursions of (146) provide an alternative to the classical fixed-interval smoothing algorithm. Thus, putting RHS of (144) and (145) into (131) and (132) respectively gives

$$\begin{aligned}b_{t|T} &= b_t + P_t q_{t+1},\\ P_{t|T} &= P_t + P_t Q_{t+1} P_t.\end{aligned} \tag{147}$$

This algorithm is due to de Jong (1989), albeit that he originally proposed to run the recursions in the opposite direction.

An alternative route to obtaining the smoothed estimates of the state vectors, which is followed by Koopman (1993), begins with the state transition equation of (53). Taking expectations, conditional upon all of the data in the sample, gives

$$E(\beta_t|\mathcal{I}_T) = \Phi_t E(\beta_{t-1}|\mathcal{I}_T) + E(\nu_t|\mathcal{I}_T). \tag{148}$$

Here, $E(\beta_{t-1}|\mathcal{I}_T) = b_{t-1|T}$ is an estimate which is assumed to have been generated already. Therefore, the task is to evaluate

$$E(\nu_t|\mathcal{I}_T) = \sum_{j=t}^{T} C(\nu_t, e_j)D^{-1}(e_j)e_j. \tag{149}$$

To find the generic covariance term $C(\nu_t, e_j)$, the recursion of (133) is run from $j-1$ down to $t-1$, and the result is substituted into (132). The only term in the resulting equation that is correlated with v_j is $H_j\Phi_j\Lambda_{j-1,t+1}(I - K_tH_t)\nu_j$. It follows that

$$\begin{aligned} C(\nu_t, e_j) &= E(\nu_t\nu_t')(I - K_tH_t)'\Lambda_{j-1,t+1}'\Phi_j'H_j', \\ &= \Psi_t(I - K_tH_t)'\Lambda_{j-1,t+1}'\Phi_j'H_j'. \end{aligned} \tag{150}$$

Putting this back in (149) gives

$$\begin{aligned} E(\nu_t|\mathcal{I}_T) &= \Psi_t(I - K_tH_t)'\sum_{j=t}^{T}\Lambda_{j-1,t+1}'\Phi_j'H_j'F_j^{-1}e_j \\ &= \Psi_t(I - K_tH_t)'q_{t+1}. \end{aligned} \tag{151}$$

This is the smoothed estimate of the state disturbance. The smoothed estimate of the state vector, which comes directly from (148), is

$$b_{t|T} = \Phi_t b_{t-1|T} + \Psi_t(I - K_tH_t)'q_{t+1}. \tag{152}$$

The initial value is $b_{0|T} = b_0 + P_0q_1$, which is obtained by setting $t = 0$ in the formula for $b_{t|T}$ under (147).

To implement the method, one must first calculate e_t, F_t^{-1} and K_t for all t via the Kalman filter that runs forward through the sample. Then the values of q_t are generated by a backwards recursion and committed to memory. Finally, the forward recursion of (152) is used in generating the smoothed disturbances and the smoothed state estimates.

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, which have been conveyed in a literature which is challenging at the best of times.

One may expect that, eventually, when these developments have been assimilated into the mainstream of econometric methodology, some of the algebraic elaborations that have accompanied them will fall into abeyance. This paper has been motivated, partly, by the thought that such a process might be hastened by assembling much of the algebra in one place in a way which demonstrates its coherence.

References

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

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

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.

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.

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.

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

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

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

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.

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

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

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.

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

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.

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.

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

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

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

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

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

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

de Jong, P., and SingFat Chu-Chun-Lin, (2002), Smoothing with an Unknown Initial Condition, Forthcoming in *The Journal of Time Series Analysis.*

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

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

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

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.

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.

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

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

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.

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

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.

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

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

Goméz, V., and A. Maravall, (1994b), Estimation, Prediction and Interpolation for Nonstationary Series with the Kalman Filter, *Journal of the American Statistical Association,* 89, 611–624.

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

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

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

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.

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.

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

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

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

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

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.

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

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.

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

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

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

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

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

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

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

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

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

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.

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

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.

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

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.

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

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

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

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

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

Pollock, D.S.G., (2002), Improved Frequency-Selective Filters, forthcoming in *Computational Statistics and Data Analysis.*

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

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.

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

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

Snyder, R.D., (1988), Computational Aspects of Kalman Filtering with a Diffuse Prior Distribution, *Journal of Statistical Computation and Simulation,* 29, 77–86.

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

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

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

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

Wellstead, P.E., and M.B. Zarrop, (1991), *Self-tuning Systems: Control and Signal Processing,* John Wiley and Sons, Chichester.

Young, P., (1984), *Recursive Estimation and Time-Series Analysis,* Springer Verlag, Berlin.

Zarrop, M.B., (1983), Variable Forgetting Factors in Parameter Estimation, *Automatica,* 19, 295–298.

Appendix

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

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

whence

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

These results are confirmed by direct multiplication.

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

$$(A.3)\qquad \begin{aligned} &\text{(i)} & (C - B'A^{-1}B)^{-1} &= C^{-1} + C^{-1}B'(A - BC^{-1}B')^{-1}BC^{-1}, \\ &\text{(ii)} & (A - BC^{-1}B')^{-1} &= A^{-1} + A^{-1}B(C - B'A^{-1}B)^{-1}B'A^{-1}, \\ &\text{(iii)} & (C + B'A^{-1}B)^{-1} &= C^{-1} - C^{-1}B'(A + BC^{-1}B')^{-1}BC^{-1}. \end{aligned}$$

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

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

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

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

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

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

where

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

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:

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

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 $B' = C(y,x)D^{-1}(x) = \Sigma_{12}\Sigma_{22}^{-1}$.

Queen Mary
University of London

This working paper has been produced by
the Department of Economics at
Queen Mary, University of London

Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096 or Fax: +44 (0)20 8983 3580
Email: j.conner@qmul.ac.uk
Website: www.econ.qmul.ac.uk/papers/wp.htm