Annals of Economic and Social Measurement, 2/4, 1973

KALMAN FILTER MODELS
A BAYESIAN APPROACH TO ESTIMATION OF TIME-VARYING REGRESSION COEFFICIENTS

BY ALEXANDER H. SARRIS*

The origins of time-varying linear regression coefficients are discussed, and it is noted that time variation cannot be estimated unless some restrictions are placed on the infinite forms of possible time changes. For example, a Markov structure imposed a priori on the coefficients renders them estimable. The structure imposes an incompletely specified prior probability distribution on the coefficients. The prior becomes completely determined through fitting it to the data. Bayes' theorem is then used to derive an estimator of the parameters. Under the assumption of perfect prior fit, the Bayes estimator is unbiased, minimum variance, and orthogonal to the residuals. Under the assumption of incomplete prior fit, the optimality properties of the estimator hold asymptotically. Finally, the problem of identifying the best Markov structure that fits the parameters is examined, and a Bayesian solution is proposed. This last discussion indicates the limitations of any method that attempts to identify time-varying coefficients.

1. INTRODUCTION

Over the last two decades great effort has been spent by econometricians, statisticians and system theorists on the problem of system identification. This problem is concerned with construction of a model whose output is close in some sense to the observed data from the real system. The modeler is guided by experience, knowledge of the real thing he is trying to describe, and intuition in specifying some equations (dynamic or static) which he terms the "structure" of the model. The equations are usually specified to within a number of parameters or coefficients which must be estimated by fitting the equations to the available data. The unknown parameters are usually assumed *a priori* to be constant. Then the problem of system identification is reduced to one of constant parameter estimation. There is a wealth of methods for the solution of this problem. A good survey of the ones that have been developed by econometricians and statisticians can be found in Theil (1971), while Åström and Eykhoff (1971) have surveyed the methods that have been developed primarily in system theory.

There are several reasons for suspecting that the parameters of many models, constructed by both engineers and econometricians, are not constant but in fact time varying. In engineering the origins of parameter variation are usually not very hard to pinpoint. Component wear, metal fatigue or component failure are some very common reasons for parameter variations. The major objective of construction of engineering models is control and regulation of the real system

* Research supported by National Science Foundation Grant GJ-1154x to the National Bureau of Economic Research, Inc.; and by an M.I.T. endowed fellowship. This Research Report, W0013 has not undergone the full critical review accorded the National Bureau's studies, including review by the Board of Directors.

I wish to acknowledge most valuable discussions with, and feedback from, Dr. Paul W. Holland of the NBER Computer Research Center, and Dr. J. Phillip Cooper of the University of Chicago. I retain sole responsibility for errors and omissions.

modeled. Therefore, most of the research in that area has concentrated on devising ways to make the output of the model insensitive to parameter variations.

On the other hand the origins of time varying parameters in econometric models are not very easy to isolate. Suspicions that shocks in the economy lead to sometimes permanent changes in the parameters of econometric models, have been substantiated ever since it was noticed that models of the economy fitted with prewar data gave noticeably different parameters than when fitted with postwar data. However, if one examines the process of economic modeling he will see several other sources of parameter variation. I will mention four of the most common ones.

Many econometric equations are mis-specified in the sense that they exclude variables that could possibly be part of the equation. Consider an equation of the form

$$y_t = \sum_{i=1}^{k} \beta_i x_{it} + \sum_{j=1}^{v} \gamma_j z_{jt} + \varepsilon_t \tag{1}$$

where y is an endogenous variable and the x_i, z_j are the true explanatory variables. If the econometrician ignores the z_j and lumps them with the error term ε, then whenever the z_j's behave in a non-stationary fashion there will be time variations in the intercept of (1).

Nonlinearities also give rise to parameter variations. If, for instance, the true relation is:

$$y_t = \alpha_1 + \alpha_2 x_t + \alpha_3 x_t^2 + \varepsilon_t \tag{2}$$

and the analyst considers the linear relation

$$y_t = \beta_1 + \beta_{2t} x_t + \varepsilon_t \tag{3}$$

then

$$\frac{\partial y_t}{\partial x_t} = \beta_{2t} = \alpha_2 + 2\alpha_3 x_t \tag{4}$$

thus β_{2t} is not constant.

Finally proxy variables and aggregation are also sources of parameter variation. For a detailed exposition of the sources of parameter variation the reader is referred to Cooley (1971).

This paper is concerned with a Bayesian method of estimation of time varying parameters. In section 2 a survey of previous research is given. The problem posed here is described in section 3. In sections 4 through 6 the method proposed for parameter estimation is presented and the properties of the estimator analysed. Sections 7 and 8 consider some problems that arise in applying the estimation technique. In section 9 the question of identifiability of a particular Markov structure is taken up, and a Bayesian solution which is the only feasible one is proposed. The last section summarizes the results.

2. Previous Research on Estimation of Time Varying Parameters

The problem of estimation of time varying parameters has not received very much attention from econometricians. On the other hand system theorists have devoted many years of research to various aspects of it. The reasons for this apparent gap will become clearer later.

The model from this point on will be assumed to be linear in the parameters. The following three classes of non-constant parameters are distinguished

(a) Time varying but non-stochastic

(b) Random but stationary

(c) Random but not necessarily stationary.

The earliest time varying parameter in econometrics dealt with parameters that were piecewise constant (Quandt (1958, 1960)) namely in class (a). This work was continued later by McGee and Carleton (1970), Brown and Durbin (1971) and Belsley (1973) but is still far from solved.

The second class of varying coefficients mentioned above applies to many problems in econometrics and statistics, and especially to the analysis of cross-sectional data. The problem is usually posed in terms of a relation of the form

$$y_t = \sum_{i=1}^{k} \beta_{it} x_{it} + \varepsilon_t \tag{5}$$

where at each period t the parameters β_{it} $(i = 1, \ldots, k)$ are a sample from a multivariate distribution with mean μ and covariance matrix Σ. The objective is usually to estimate μ and Σ. Work on this problem has been done by Rao (1965), Hildreth and Houck (1968), Burnett and Guthrie (1970), Swamy (1970), and Rosenberg (1972).

Under the third category mentioned above come the various sequential variation models of the form

$$\beta_{t+1} = T\beta_t + u_t. \tag{6}$$

This model is very common in the engineering literature and can be utilized to represent a wide variety of sample paths. In the econometrics literature to my knowledge only Rosenberg (1967, 1968a, b) has dealt extensively with this kind of sequential variation. Cooley (1971) has also used it, mainly as a predictive tool.

On the other hand the engineering literature on estimation of models of the form (6) is huge. The earliest work was the one by Kalman and Bucy (1961). For extensive bibliographies and various aspects of the problem the reader can consult the textbooks of Sage and Melsa (1971), and Åström (1970) as well as the special issue of the IEEE (1971) Transactions on Automatic Control.

In most of the engineering literature the statistics of the uncertain quantities are assumed known. This is a severe restriction when one is transferred to the realm of statistics and econometrics and is one of the primary reasons for which there is a large gap between research in system theory and the quantitative social sciences. Interesting exceptions to the rule in the engineering literature are the papers by Mehra (1970, 1971, 1972), and Kashyap (1970). Furthermore, the engineers usually make strong *a priori* assumptions about the matrix T, which as will be seen in section 9 do not, in general, hold in an econometric framework.

3. Problem Description

Consider the following model

$$y_t = x_t\beta_t + \varepsilon_t \tag{7}$$

where y_t is the response to the effects of the k explanatory variables $x_{1t}, x_{2t}, \ldots, x_{kt}$, x_t is a $1 \times k$ vector of the mentioned explanatory variables, β_t is a $k \times 1$ vector of time varying coefficients, and ε_t is a disturbance term that is assumed to be normally distributed with the following properties.

$$E[\varepsilon_t] = 0 \tag{8}$$

$$E[\varepsilon_t^2] = \sigma_\varepsilon^2 \tag{9}$$

$$E[\varepsilon_t \varepsilon_l] = \sigma_\varepsilon^2 \delta_{kl} \tag{10}$$

$$E[\varepsilon_t \beta_l] = 0 \tag{11}$$

where δ_{kl} is the Kronecker delta, and σ_ε^2 is an unknown constant. The assumption is that there are N observations on the endogenous variable y and the k exogenous variables.

Define the following quantities

$$y = [y_1, y_2, \ldots, y_N]' \tag{12}$$

where (′) denotes the transposition.

$$\beta = [\beta'_1, \beta'_2, \ldots, \beta'_N]' \tag{13}$$

$$\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N]' \tag{14}$$

$$X = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & & & x_N \end{bmatrix}. \tag{15}$$

The available information now can be written as follows:

$$y = X\beta + \varepsilon. \tag{16}$$

It can be readily seen now that it is impossible to estimate the vector β (a $Nk \times 1$ vector) from (16), via ordinary least squares (OLS) regression. To use the OLS formula the matrix $X'X$ must be invertible. It is easily seen, however, that this $Nk \times Nk$ matrix has rank at most equal to N. So there are not enough degrees of freedom to estimate β.

The conclusion from the above discussion is that there is no hope of estimating β unless some more information about the vector becomes available. I will assume that the β_i's can be generated by a Markovian structure of the form

$$\beta_{t+1} = T\beta_t + u_{t+1} \qquad (t = 0, 1, \ldots, N-1) \tag{17}$$

where: T is a known $k \times k$ transition matrix and u_j is a $k \times 1$ vector of random shocks distributed as multivariate normal with zero mean and covariance matrix

$$E[u_j u'_k] = \sigma_u^2 R \delta_{kj} \tag{18}$$

where R is a known $k \times k$ positive semidefinite matrix.

The vector β_0 will be assumed unknown.

With the assumption of a structure such as the above, what is achieved is a prior distribution on the vector β. However, this distribution is not completely known because σ_u^2 and β_0 are not known. So it is not a complete Bayesian prior and so Bayesian analysis cannot be carried out immediately. Some "fitting" must be done before the Bayesian analysis is started.

The idea of an incomplete prior distribution might seem strange. A rationalization of it is the following. With the assumption of (17) the vector valued time series $\{\beta_t\}$ is restricted to a class of particular sample paths. However, the econometrician is ignorant about the level at which the sample paths start, and about the spread that he can allow the class of sample paths to have. He expresses this by letting the data define these quantities for him.

The problem now is two-fold. First find the prior for β that best fits the data. Then use the calculated prior to carry out a prior to posterior analysis to obtain the posterior distribution of β.

4. Estimation of the Best Prior of β

In this section maximum likelihood is used to estimate the quantities β_0, σ_ε^2 and σ_u^2. First β_k is expressed in terms of β_0.

(19)
$$\begin{aligned}\beta_k &= T\beta_{k-1} + u_k = T(T\beta_{k-2} + u_{k-1}) + u_k \\ &= T^2\beta_{k-2} + u_k + Tu_{k-1} = \dots \\ &= T^k\beta_0 + \sum_{j=1}^{k} T^{k-j}u_j.\end{aligned}$$

By substituting in (7) the following is obtained

(20)
$$y_k = x_k\beta_k + \varepsilon_k = x_kT^k\beta_0 + x_k \sum_{j=1}^{k} T^{k-j}u_j + \varepsilon_k.$$

Define:

(21)
$$z_k \equiv x_kT^k \qquad (k = 1, 2, \dots, N)$$

(22)
$$v_k \equiv x_k \sum_{j=1}^{k} T^{k-j}u_j + \varepsilon_k \qquad (k = 1, 2, \dots, N).$$

Letting

(23)
$$Z \equiv [z'_1, z'_2, \dots, z'_N]'$$

and

(24)
$$v \equiv [v_1, v_2, \dots, v_N]'$$

the relations (20) can be written compactly as follows:

(25)
$$y = Z\beta_0 + v.$$

The vector v is distributed as multivariate normal with mean

$$E(v) = 0$$

and covariance matrix given by the following equations:

$$Ev_k^2 = E\left[x_k \sum_{j=1}^{k} T^{k-j}u_j + \varepsilon_k\right]\left[\varepsilon_k + \left(\sum_{j=1}^{k} u_j' T'^{k-j}\right)x_k'\right]$$

(26) $$= \sigma_\varepsilon^2 + \sigma_u^2 x_k\left(\sum_{j=1}^{k} T^{k-j}RT'^{k-j}\right)x_k'$$

$$Ev_kv_l = E\left[x_k \sum_{j=1}^{k} T^{k-j}u_j + \varepsilon_k\right]\left[\varepsilon_l + \left(\sum_{j=1}^{l} u_j' T'^{l-j}\right)x_l'\right]$$

(27) $$= x_k\left(\sum_{j=1}^{l} T^{k-j}RT'^{l-j}\right)x_l' \qquad \text{if } k > l$$

(28) $$= x_k\left(\sum_{j=1}^{k} T^{k-j}RT'^{l-j}\right)x_l' \qquad \text{if } k < l.$$

It is readily noticed that (27) and (28) give the same quantities. The covariance of v can be written as

(29) $$(Evv') = \sigma_\varepsilon^2 I + \sigma_u^2 Q$$

where I is a $N \times N$ unit matrix and Q is a $N \times N$ known matrix with entries

(30) $$Q_{ij} = x_i\left(\sum_{n=1}^{\min(i,j)} T^{i-n}RT'^{j-n}\right)x_j' = Q_{ji}.$$

Furthermore define

(31) $$\theta \equiv \sigma_u^2/\sigma_\varepsilon^2$$

and write

(32) $$E(vv') = \sigma_\varepsilon^2(I + \theta Q) = \sigma_\varepsilon^2 P(\theta).$$

So the covariance matrix of v is known up to two scalar constants.

The logarithm of the likelihood of y can now be written as

(33) $$L(y; Z, \beta_0, \sigma_\varepsilon^2, \theta) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma_\varepsilon^2 - \tfrac{1}{2}\ln|P(\theta)| - \frac{1}{2\sigma_\varepsilon^2}(y - Z\beta_0)'P(\theta)^{-1}(y - Z\beta_0)$$

where $|P|$ denotes the determinant of P. The values of β_0 and σ_ε^2 that maximize this likelihood are

(34) $$\hat{\beta}_0(\theta) = [Z'P^{-1}(\theta)Z]^{-1}Z'P^{-1}(\theta)y$$

(35) $$\hat{\sigma}_\varepsilon^2(\theta) = \frac{[y - Z\hat{\beta}_0(\theta)]'P^{-1}(\theta)[y - Z\hat{\beta}_0(\theta)]}{N}$$

(36) $$= \frac{y'P(\theta)^{-1}[I - Z[Z'P(\theta)^{-1}Z]^{-1}Z'P(\theta)^{-1}]y}{N}.$$

Now by substituting the expressions for $\hat{\beta}_0(\theta)$ and $\sigma_\varepsilon^2(\theta)$ in the logarithmic likelihood function, the concentrated likelihood for θ is obtained.

$$(37) \qquad L(y; Z, \theta) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\frac{1}{N}\{y'P^{-1}[I - Z(Z'P^{-1}Z)^{-1}Z'P^{-1}]y\} - \frac{1}{2}\ln|P(\theta)| - \frac{N}{2}.$$

The concentrated likelihood function can be maximized numerically for θ to obtain a maximum likelihood estimator of θ and the procedure is over. The only possible difficulty is the inversion of the $N \times N$ matrix $P(\theta)$ for every value of θ. This can be avoided, however, with the following trick.

$$P(\theta) = I + \theta Q.$$

Find an orthogonal matrix G which diagonalizes Q, so that

$$(38) \qquad Q = G'DG$$

with $G'G = I$ and D a $N \times N$ diagonal matrix. Then

$$(39) \qquad P(\theta) = I + \theta G'DG = G'[I + \theta D]G = G'D^*(\theta)G$$

where $D^*(\theta)$ is diagonal. Hence

$$(40) \qquad P^{-1}(\theta) = G'D^{*-1}(\theta)G$$

and inversion of $P^{-1}(\theta)$ is reduced to inversion of a diagonal matrix which is trivial.

Note that if y^* and Z^* are defined as

$$(41) \qquad y^* = Gy$$

$$(42) \qquad Z^* = GZ$$

the estimators for β_0 and σ_ε^2 become

$$(43) \qquad \hat{\beta}_0(\theta) = [Z^{*\prime}D^{*-1}(\theta)Z^*]^{-1}Z^{*\prime}D^{*-1}(\theta)y^*$$

$$(44) \qquad \hat{\sigma}_\varepsilon^2(\theta) = \frac{1}{N}\{y^{*\prime}D^{*-1}(\theta)[I - Z^*[Z^{*\prime}D^{*-1}(\theta)Z^*]^{-1}Z^{*\prime}D^{*-1}(\theta)]y^*\}$$

and they are equal to the ones given by (34)–(36).

Note that the diagonalization is to be done only once and not at every iteration for θ. This is a significant computational advantage. There exist very efficient algorithms for achieving diagonalization for large non-sparse matrices. I have heard of a procedure at Argonne National Laboratories that took 80 seconds on a IBM-195 computer to diagonalize a 768 × 768 dense matrix.[1]

The procedure outlined above for finding the best prior is not new. Cooley (1971) has used a similar procedure, although his problem was quite different than

[1] I owe some of these comments to Dr. Virginia Klema of the N.B.E.R. Computer Research Center.

mine. He has given complete proofs of the consistency and efficiency of the estimators obtained by this procedure and the interested reader is referred to his work.

Note that the estimator for β_0 is unbiased and identical to the Generalized Least Squares (GLS) estimator obtained by minimizing

$$(45) \qquad (y - Z\beta_0)'[P(\theta)]^{-1}(y - Z\beta_0).$$

The estimator of σ_ε^2 is biased because

$$(46) \qquad E\hat{\sigma}_\varepsilon^2(\theta) = \frac{N-k}{N}\sigma_\varepsilon^2.$$

For large N the bias is negligible. The imbiased estimator of σ_ε^2

$$_u\hat{\sigma}_\varepsilon^2 \equiv \frac{N}{N-k}\hat{\sigma}_\varepsilon^2(\theta)$$

which is obtained from GLS could also be used instead of the one given in (35). Presumably, the estimate of θ would be slightly different, but for large N the difference would be minor.

Another approach that could have been used, is to assess a prior distribution jointly for β_0, σ_ε^2 and θ. Then Bayes' theorem could be used to estimate the posterior distribution of β_0, σ_ε^2 and θ, and inferences about the unknown quantities could thus be made.

It seems to the author, however, that the econometrician will almost never have any prior information or feeling about the above mentioned unknown quantities. Assuming diffuse priors on the other hand would not lead to any substantially different results than the ones obtained by the maximum likelihood approach mentioned.

5. Bayesian Estimation of the Time Varying Coefficients

In the previous section the prior distribution of the $Nk \times 1$ vector β was assessed. For the purposes of this and the next section the constants β_0, σ_ε^2 and θ will be assumed known accurately. Consider now the model (16)

$$y = X\beta + \varepsilon$$

β has a multivariate normal prior density with mean

$$(48) \qquad \mu = E(\beta) = \begin{bmatrix} E(\beta_1) \\ E(\beta_2) \\ \vdots \\ E(\beta_N) \end{bmatrix} = \begin{bmatrix} T\beta_0 \\ T^2\beta_0 \\ \vdots \\ T^N\beta_0 \end{bmatrix},$$

and covariance matrix

$$(49) \qquad E[(\beta - \mu)(\beta - \mu)'] \equiv V$$

where
$$V_{ij} \equiv \operatorname{cov}(\beta_i, \beta_j) = E\{[\beta_i - E(\beta_i)][\beta_j - E(\beta_j)]'\}$$
$$= E\left\{\left[\sum_{n=1}^{i} T^{i-n}u_n\right]\left[\sum_{m=1}^{j} T^{j-m}u_m\right]'\right\}$$
$$= \sigma_u^2 \sum_{n=1}^{\min(i,j)} T^{i-n}RT'^{j-n}.$$

If the following matrix is defined

(51)
$$M = \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ T & I & 0 & \dots & 0 \\ T^2 & T & I & 0 & \dots \\ \vdots & & & & \\ T^{N-1} & \dots & \dots & \dots & I \end{bmatrix}$$

then

(52)
$$V = \sigma_u^2 M(R \otimes I_N)M'$$

where I_N is the $N \times N$ unit matrix, and $\otimes$ denotes the Kronecker product of two matrices. Note that V is invertible if and only if R is invertible. It will be assumed that R is positive definite so that V is invertible.

The likelihood of the data y given β is multivariate normal with mean $X\beta$ and variance $\sigma_\varepsilon^2 I$. The joint density of y and β can thus be found by multiplying the prior of β and the likelihood function

(53)
$$p(y, \beta) = p(\beta)l(y; X, \beta) = \frac{1}{(2\pi)^{Nk/2}|V|^{1/2}} \exp\{-\tfrac{1}{2}(\beta - \mu)'V^{-1}(\beta - \mu)\}$$
$$\cdot \frac{1}{(2\pi)^{N/2}\sigma_\varepsilon^N} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(y - X\beta)'(y - X\beta)\right\}.$$

The quadratic form of β in the exponent is now manipulated so as to complete the square in β.

(54)
$$\frac{1}{\sigma_\varepsilon^2}(y - X\beta)'(y - X\beta) + (\beta - \mu)'V^{-1}(\beta - \mu)$$
$$= \left[\beta - \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1}\left(V^{-1} + \frac{1}{\sigma_\varepsilon^2}X'y\right)\right]'\left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)$$
$$\cdot \left[\beta - \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1}\left(V^{-1} + \frac{1}{\sigma_\varepsilon^2}X'y\right)\right]$$
$$+ \left\{\frac{1}{\sigma_\varepsilon^2}y'y + \mu'V^{-1}\mu - \left(V^{-1}\mu + \frac{1}{\sigma_\varepsilon^2}X'y\right)'\right.$$
$$\left.\cdot \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1}\left(V^{-1}\mu + \frac{1}{\sigma_\varepsilon^2}X'y\right)\right\}.$$

The joint density thus becomes easily integrable with respect to β. After β is integrated out to obtain the marginal density of y, the posterior density of β given y is obtained as following

$$(55)\quad p(\beta|y) = \frac{p(y,\beta)}{p(y)} = \frac{|(1/\sigma_\varepsilon^2)X'X + V^{-1}|}{(2\pi)^{Tk/2}} \exp\left\{-\frac{1}{2}\left[\beta - \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1} \cdot \left(V^{-1}\mu + \frac{1}{\sigma_\varepsilon^2}X'y\right)\right]'\left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)\left[\beta - [\cdot]^*\right]\right\}.$$

The posterior density is therefore multivariate normal with mean

$$(56)\quad \hat{\beta} = E(\beta|y) = \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1}\left(V^{-1}\mu + \frac{1}{\sigma_\varepsilon^2}X'y\right)$$

and variance

$$(57)\quad E\{[\beta - E(\hat{\beta}|y)][\beta - E(\hat{\beta}|y)]|y\} = \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1}.$$

If the loss used in calculating the best estimator is quadratic, namely of the form

$$E[\hat{\beta} - \beta)'(\hat{\beta} - \beta)]$$

then it is well known that the best estimator is the mean of the posterior density of β. Therefore, for quadratic loss equation (56) gives the optimal estimator. Now equation (56) involves inversion of a $Nk \times Nk$ matrix. This is excessive. To circumvent the problem the matrix inversion lemma which is stated here without proof is invoked. (For a proof see e.g. Duncan and Horn (1972).)

Lemma 5.1. If $S = [M^{-1} + AR^{-1}B]^{-1}$ then

$$(58)\quad S = M - MA[R + BMA]^{-1}BM.$$

Using (58) the Nk $\times$ Nk matrix in (56) can be written

$$(59)\quad \left(\frac{1}{\sigma_\varepsilon^2}X'X + V^{-1}\right)^{-1} = V - VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV$$

Theorem 5.1. Suppose that the prior density of β is multivariate normal with mean μ and covariance matrix V. Suppose that V is positive definite. If the likelihood of the data y given β is multivariate normal with mean $X\beta$ and covariance matrix $\sigma_\varepsilon^2 I_N$, then the posterior density of β given the data y is multivariate normal with mean

$$(60)\quad E(\beta|y) = \mu + VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}(y - X\mu)$$

and covariance matrix equal to

$$(61)\quad CB = V - VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV.$$

* $[\cdot]$ denotes the expression for the mean of β appearing in the first bracket.

Proof. This theorem summarizes the results proved above. The only thing that needs proof is equation (60). (56) with the aid of lemma 5.1 becomes

$$E(\beta|y) = [V - VX'(\sigma_\varepsilon^2 I_N + XVX')^{-1}XV]\left(V^{-1}\mu + \frac{1}{\sigma_\varepsilon^2}X'y\right)$$

$$= \mu + VX'\left[\frac{1}{\sigma_\varepsilon^2}y - (\sigma_\varepsilon^2 I_N + XVX')^{-1}X\mu - \frac{1}{\sigma_\varepsilon^2}(\sigma_\varepsilon^2 I_N + XVX')^{-1}XVX'y\right]$$

$$= \mu + VX'(\sigma_\varepsilon^2 I_N + XVX')^{-1}\left[(\sigma_\varepsilon^2 I_N + XVX')\frac{1}{\sigma_\varepsilon^2}y - X\mu - \frac{1}{\sigma_\varepsilon^2}XVX'y\right]$$

$$= \mu + VX'(\sigma_\varepsilon^2 I_N + XVX')^{-1}(y - X\mu).$$ ■

It is of considerable interest to notice that the matrix $\sigma_\varepsilon^2 I_N + XVX'$ is nothing but the matrix $\sigma_\varepsilon^2 P(\theta)$ defined by (32), and whose elements are given by equations (26)–(28). Since the inversion of this matrix has been accomplished during the first part of the estimation process, namely the "fitting" part, formula (60) provides an easy way to estimate the complete series $(\beta_1, \ldots, \beta_N)$ at once. A simplification of equation (60) is now given, that will enhance the reader's intuition about (60), and will clarify the "smoothing" character of the estimator.

Theorem 5.2. The Bayesian estimator of β is equivalent to the following sequential estimator

(62) $$\hat{\beta}_{i|N} = T\hat{\beta}_{i-1|N} + \frac{1}{\sigma_\varepsilon^2}RM_iP^{-1}(\theta)(y - X\mu)$$

where $\hat{\beta}_{0|N} = \beta_0$,

(63) $$M_i = [0 \quad 0, \ldots, x_i', T'x_{i+1}', \ldots, T'^{N-i}x_N']$$

a $k \times N$ matrix, and $\beta_{i|N}$ denotes the ith $(k \times 1)$ vector component of $\hat{\beta}$ (cf. (13)).

Proof. The proof hinges on observing the structure of the matrix V. Denote by V_1 the first k rows of V, by V_2 the next k rows, etc. up to V_N. It is then easy to see, having in mind the definition of V by (50) that

(64) $$V_i = TV_{i-1} + F_i$$

where

(65) $$F_i = [0 \quad 0 \ldots R, RT', \ldots, RT'^{N-1}]$$

then

(66) $$V_iX' = TV_{i-1}X' + F_iX' = TV_{i-1}X' + RM_i$$

Defining $\mu_i = T^i\beta_0$ and using (60)

$$\begin{aligned}
\hat{\beta}_{i|N} &= \mu_i + V_iX'\left(\frac{1}{\sigma_\varepsilon^2}[P(\theta)]^{-1}\right)(y - X\mu) \\
&= T\mu_{i-1} + [TV_{i-1}X' + RM_i]\left(\frac{1}{\sigma_\varepsilon^2}[P(\theta)]^{-1}\right)(y - X\mu) \\
&= T\left[\mu_{i-1} + V_{i-1}X'\left(\frac{1}{\sigma_\varepsilon^2}[P(\theta)]^{-1}\right)(y - X\mu)\right] \\
&\quad + RM_i\left(\frac{1}{\sigma_\varepsilon^2}[P(\theta)]^{-1}\right)(y - X\mu) \\
&= T\hat{\beta}_{i-1|N} + \frac{1}{\sigma_\varepsilon^2}RM_i[P(\theta)]^{-1}(y - X\mu). \qquad \blacksquare
\end{aligned}$$

Theorem 5.2 shows explicitly how data subsequent to period i enter the estimation process. Notice also that the covariance matrix given in (61) is more general than the one traditionally deduced in the engineering literature. There, interest centers mostly on the covariance matrices of $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_N$ and not on the cross-covariance matrices between $\hat{\beta}_i$ and $\hat{\beta}_j$ for $i \neq j$. Equation (61) gives explicitly all the necessary covariances. Equation (62) is a so-called "smoothing" equation, because it shows the effect on β_i of observations obtained before as well as after time i.

As far as the author is aware, the combined Bayesian estimation procedure for all the unknown parameters presented in sections 4 and 5, has not appeared in the literature before.

6. Properties of the Bayesian Estimator

In this section some small and large sample properties of the estimator obtained in the previous section are examined.

Theorem 6.1. (Unbiasedness.) Suppose that the explanatory variables x_t do not contain any lagged values of the endogenous variable y. Then the Bayesian estimator given by (60) is unbiased, in the sense that $E(\hat{\beta} - \beta) = 0$.

Proof. If there are no lagged values of y in x, then the matrix $\sigma_\varepsilon^2 I_N + XVX'$ is not a random variable. Therefore:

$$\begin{aligned}
E(\hat{\beta} - \beta) &= E\{\mu + VX'(\sigma_\varepsilon^2 I_N + XVX')^{-1}(y - X\mu) - \beta\} \\
&= \mu + VX'(\sigma_\varepsilon^2 I_N + XVX')^{-1}[E(y) - X\mu] - E\beta \\
&= \mu + VX'(\sigma_\varepsilon^2 I_N + XVX')^{-1}(X\mu - X\mu) - \mu = 0. \qquad \blacksquare
\end{aligned}$$

If the explanatory variables contain lagged values of the endogenous variable, then the estimator obtained in the previous section remains unchanged because the likelihood function is the same. This can be seen since

$$l(y;\beta) = p(y|\beta) = p(y_1|\beta)p(y_2|y_1;\beta)\ldots p(y_N|y_1, y_2, \ldots, y_{N-1};\beta). \tag{67}$$

However, the estimator is now biased in a complicated way. Some experimental results employing a special case of the estimator (60) reported by Wieslander and Wittenmark (1971) support the hypothesis that the estimator in this case has a small negative bias.

For the rest of this section the assumption is that there are no lagged values of the endogenous variable in X.

Theorem 6.2. (Properties of $\hat{\beta}$.) The Bayesian estimator of β obtained in the previous section and given in theorem 5.1 has the following properties:

(1) $\hat{\beta} - \beta$ is distributed as multivariate normal with mean equal to zero and variance $V - VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV$.

(2) $E(\hat{\beta} - \beta)(y - X\beta)' = 0$.

(3) $E(\hat{\beta} - \beta)y' = 0$.

Proof.

(1) It was seen in theorem 6.1 that $\hat{\beta}$ has mean μ. Its variance is

$$E[(\hat{\beta} - \mu)(\hat{\beta} - \mu)'] = VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}E[(y - X\mu)(y - X\mu)'] \cdot [\sigma_\varepsilon^2 I_N + XVX']^{-1}XV \tag{68}$$

$$\begin{aligned} E[(y - X\mu)(y - X\mu)'] &= E[(X(\beta - \mu) + \varepsilon)(X(\beta - \mu) + \varepsilon)'] \\ &= XVX' + \sigma_\varepsilon^2 I_N. \end{aligned} \tag{69}$$

Therefore

$$E[(\hat{\beta} - \mu)(\hat{\beta} - \mu)'] = VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV \tag{70}$$

$$\begin{aligned} E[(\hat{\beta} - \mu)(\beta - \mu)'] &= VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}E[(y - X\mu)(\beta - \mu)'] \\ &= VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}E\{[X(\beta - \mu) + \varepsilon](\beta - \mu)'\} \\ &= VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV. \end{aligned} \tag{71}$$

Now using (70) and (71)

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] &= E\{[(\hat{\beta} - \mu) - (\beta - \mu)][(\hat{\beta} - \mu) - (\beta - \mu)]'\} \\ &= E[(\hat{\beta} - \mu)(\hat{\beta} - \mu)'] + E[(\beta - \mu)(\beta - \mu)'] \\ &\quad - E[(\hat{\beta} - \mu)(\beta - \mu)'] - E[(\beta - \mu)(\hat{\beta} - \mu)'] \\ &= V - VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV. \end{aligned}$$

$$y - X\hat{\beta} = X(\beta - \hat{\beta}) + \varepsilon \tag{72}$$

$$E[(\hat{\beta} - \beta)(y - X\beta)'] = E[-(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X' + (\hat{\beta} - \beta)\varepsilon'] \tag{73}$$

$$\begin{aligned} E[(\hat{\beta} - \beta)\varepsilon'] &= E\{[(\mu - \beta) + VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1} \\ &\quad \cdot [X(\beta - \mu) + \varepsilon]]\} \\ &= VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}\sigma_\varepsilon^2 I_N. \end{aligned} \tag{74}$$

Using property (1) and equations (73) and (74) the following is obtained

$$E[(\hat{\beta} - \beta)(y - X\hat{\beta})'] = -VX' + VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XVX'$$
$$+ VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}\sigma_\varepsilon^2 I_N = -VX' + VX' = 0.$$

(3) $$E[(\hat{\beta} - \beta)y'] = E[(\hat{\beta} - \beta)(y' - \hat{\beta}'X' + \hat{\beta}'X')]$$
$$= E[(\hat{\beta} - \beta)\hat{\beta}']X' = E\{[(\hat{\beta} - \mu) - (\beta - \mu)][\hat{\beta}' - \mu' + \mu']\}X'$$
$$= E[(\hat{\beta} - \mu)(\hat{\beta} - \mu)']X' - E[(\beta - \mu)(\hat{\beta} - \mu)']X'$$
$$+ E[(\hat{\beta} - \beta)]\mu'X'$$
$$= VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XVX' - VX'[\sigma_\varepsilon^2 I_N$$
$$+ XVX']^{-1}XVX'$$
$$= 0.$$ ■

The above theorem indicates that $\hat{\beta}$ is the projection of β on y.

Notice that $\hat{\beta}$ is affine in y; i.e. of the form $Ay + a$. The following theorem proves that $\hat{\beta}$ has minimum variance among the class of unbiased affine estimators of β.

Theorem 6.3. Let $\tilde{\beta}$ be any estimator of β that is affine in y and unbiased. Then the matrix

$$A = E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] - E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

is positive semidefinite.

Proof. Write

(75) $$\tilde{\beta} = \hat{\beta} + Hy + h.$$

Unbiasedness yields

(76) $$E(\tilde{\beta} - \beta) = E(\hat{\beta} - \beta + Hy + h) = HX\mu + h = 0$$

(77) $$E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] + E[(Hy + h)(Hy + h)']$$
$$+ E[(\hat{\beta} - \beta)(Hy + h)'] + E[(Hy + h)(\hat{\beta} - \beta)'].$$

Unbiasedness and property (3) of theorem 6.2 render the cross terms equal to zero. The matrix A now becomes

$$A = E[(Hy + h)(Hy + h')] = E(ff')$$

so indeed A is always a positive semidefinite matrix. ■

Although the estimator of β was derived using the Bayesian framework, it is of interest to examine its large sample properties. The arguments will be sketchy since a lot of discussions have appeared elsewhere. Large sample properties of posterior distributions have been examined among others by Jeffreys (1961, p. 193), Johnson (1967), and Zellner (1971, p. 31). It has been shown that under mild assumptions the posterior distribution for a vector of parameters approaches a normal distribution with mean equal to the maximum likelihood estimate, and covariance matrix equal to the information matrix. The way the maximum likelihood estimate of β_k is derived is shown below.

The first step is to express all vector β_i with $i \neq k$ in terms of β_k as following

$$\beta_i = (T^{-1})^{k-i}\beta_k - \sum_{j=1}^{k-i} (T^{-1})^{j-1}u_{i+j} \quad \text{for } i < k \tag{78}$$

$$\beta_i = \beta_k \quad \text{for } i = k \tag{79}$$

$$\beta_i = T^{i-k}\beta_k + \sum_{j=1}^{i-k} T^{i-k-j}u_{k+j} \quad \text{for } i > k. \tag{80}$$

Then the data is written with β_k appearing as the only unknown.

$$y = Z^{(k)}\beta_k + v^{(k)} \tag{81}$$

where

$$Z^{(1)} = \begin{bmatrix} x_1 \\ x_2 T \\ \vdots \\ x_N T^{N-1} \end{bmatrix} \tag{82}$$

$$Z^{(k)} = Z^{(k-1)}T^{-1} \tag{83}$$

$$Ev^{(k)}v^{(k)\prime} = \sigma_\varepsilon^2 I + \sigma_u Q^{(k)} \equiv \sigma_\varepsilon^2 P^{(k)}(\theta) \tag{84}$$

where

$$Q^{(k)} = X[N^{(k)}(R \otimes I_N)N^{(k)\prime} + M^{(k)}(R \otimes I_N)M^{(k)\prime}]X' \tag{85}$$

and

$$N^{(k)} = \begin{bmatrix} I & T^{-1} & T^{-2} & \dots & T^{-(k-2)} & 0 & \dots & 0 \\ 0 & I & T^{-1} & \dots & T^{-(k-3)} & 0 & \dots & 0 \\ \vdots & 0 & I & \ddots & & & & \\ \vdots & & & & I & & & \\ & & & & & 0 & & \\ & & & & & & \ddots & \\ 0 & & & & & & & 0 \end{bmatrix} \tag{86}$$

$$M^{(k)} = \begin{bmatrix} 0 & & & & & & \\ \vdots & \ddots & & 0 & & & \\ 0 & & & & & & \\ I & & 0 & & & & \\ T & & I & 0 & & & \\ T^2 & & T & I & & & \\ \vdots & & & & \ddots & & \\ T^{N-1-k} & & \dots & & I & 0 \; \dots & 0 \end{bmatrix} \tag{87}$$

where

$$N^{(1)} = 0$$

and

(88)
$$N^{(2)} = \begin{bmatrix} I & & \mathbf{0} \\ & 0 & \\ \mathbf{0} & & \ddots \end{bmatrix}.$$

The maximum likelihood estimator of β_k is then

(89)
$$\beta_k = [Z^{(k)\prime}[P^k(\theta)]^{-1}Z^{(k)}]^{-1}Z^{(k)\prime}[P^{(k)}(\theta)]^{-1}y.$$

Notice that for $k = 0$ (79) reduces to (34); i.e. the maximum likelihood estimator of β_0 obtained earlier. Cooley (1971) has proven that this estimator of β_k is consistent and efficient. Since for large samples the Bayesian estimate of β_k approaches the maximum likelihood estimate of β_k, it also will be consistent and efficient. However, besides this point there are obvious advantages to the Bayesian estimators since their complete distribution is known, and the interactions between β_i and β_j for $i \neq j$ are easily seen.

7. Effects of Errors in β_0, σ_ε and θ

The results of the previous two sections hold only if the parameters $\beta_0, \sigma_\varepsilon$ and θ are known with certainty. Since $\beta_0, \sigma_\varepsilon$ and θ are not known with accuracy, the errors of $\hat{\beta}$ in estimating β will be compounded. Cooley (1971) has proven in a similar context that the estimators of quantities like $\beta_0, \sigma_\varepsilon$ and θ are consistent and efficient. So the results of the previous two sections are certainly true for large samples. For small samples, experimental results would indicate the validity of the theory. The problem of time varying parameters, however, is new to econometrics and statistics and published experimental work is still lacking.

On the other hand, the engineering literature has touched on the subject with research under the general name of "adaptive filtering." Heffes (1966) has mentioned that in general the estimator obtained with erroneous parameters will not be minimum variance. However, the error will not, in general, be large. Work on devising algorithms to compensate for the errors has been reported by Mehra (1970, 1972). The subject, however, has still not come under detailed scrutiny.

8. Effects of Singular R Matrix

In section 5 the covariance matrix of the shocks that change the coefficients was assumed positive definite. This was done to guarantee the invertibility of V (i.e. its positive definiteness) and the validity of the Bayesian approach. In most practical cases, however, the case will be that R is singular. If, for example, one of the coefficients is assumed to remain unchanged then the corresponding column and row of R will be zero.

Notice that theorem 5.1 indicates a form of the estimator which requires only the inversion of $XVX' + \sigma_\varepsilon^2 I_N$. This inversion can always be done, as long as V is positive semidefinite. Equation (60) was derived through the Bayesian approach, but itself does not require nonsingularity of R. The keen reader must have noticed that in the proof of all the theorems of section 6, only form (60)–(61) of the estimator and its variance was used. The following theorem has thus been proved implicitly in the previous sections.

Theorem 8.1. Let the covariance matrix R, of the coefficient shocks, be singular. Then the minimum variance estimator of β is

$$\hat{\beta} = \mu + VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}(y - X\mu) \tag{90}$$

and has covariance matrix equal to

$$V - VX'[\sigma_\varepsilon^2 I_N + XVX']^{-1}XV. \tag{91}$$

9. The Identifiability of the Transition Matrix T

Throughout the previous sections the transition matrix T was assumed constant and known. In this section this assumption is relaxed, and the consequences of alternate T's are examined.

The imposition of a transition relation is crucial to the determination of the prior distribution of the time varying coefficients. The choice of an appropriate T reflects the analyst's prior beliefs about the class of sample paths that he will allow the β to be a member of. It is, therefore, of considerable interest to examine members of various classes of sample paths that arise from consideration of different T's. The autoregressive integrated moving average (ARIMA) class of time series models, analyzed in depth by Box and Jenkins (1970), is general enough to describe most sample paths of interest. ARIMA models are capable of generating trends, cycles, as well as violent fluctuations.

To obtain a feeling for the kinds of sample paths that ARIMA models generate, a scalar parameter β_t will be considered. First assume that the matrix T is the unit matrix, namely $T = 1$ in the scalar case considered here. Then the Markov structure imposed on the varying coefficient is of the form

$$\beta_t = \beta_{t-1} + u_t. \tag{92}$$

This is the most commonly used *a priori* Markov structure, and is appealing because of its simplicity. Figure 9.1 shows typical sample paths generated by this kind of structure (these and all the subsequent sample paths were generated by Monte Carlo simulations of the relevant structures). What is evident from the figure is that structure (92) leads to very noisy time series. If the analyst feels *a priori* that the parameters of the model are varying violently from period to period, then (92) seems an appropriate structure.

In many cases, however, the *a priori* belief might be that the parameters drift slowly across time. In these situations the following model might seem more appropriate.

$$\beta_t = 2\beta_{t-1} - \beta_{t-2} + u_t. \tag{93}$$

This model implies that the second time difference is stationary, as opposed to stationary first difference used in (92). Equation (93) can be reduced to the familiar

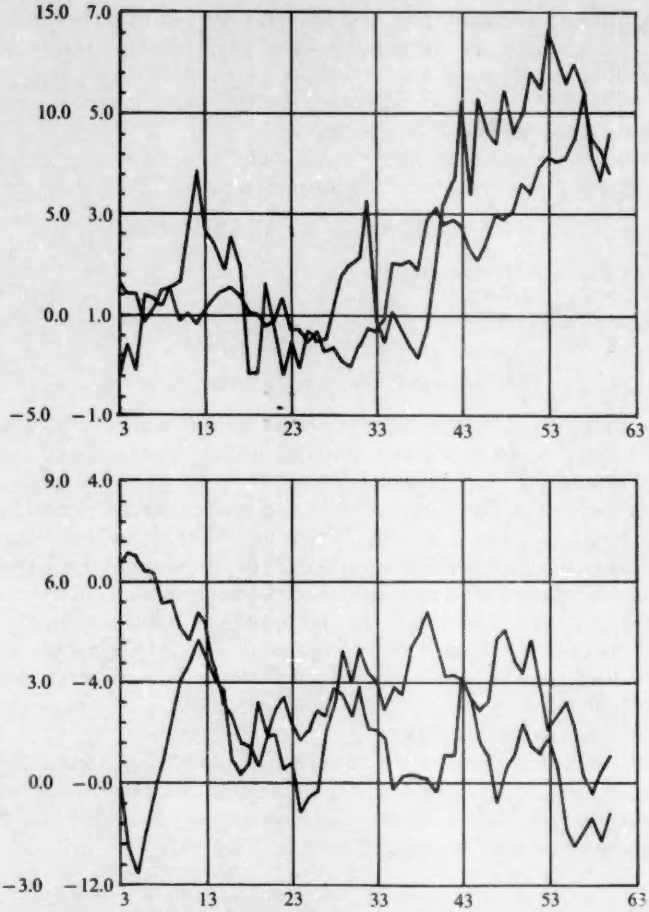

Figure 9.1 Time Series Sample Paths of the Structure $\beta_t = \beta_{t-1} + u_t$. The initial conditions are zero for all paths; u_t is normal white disturbances.

form by the definitions

(94) $$\gamma_t = \beta_{t-1}$$

(95) $$\delta_t = \beta_t.$$

Then (93) becomes

(96) $$\begin{bmatrix} \gamma_t \\ \delta_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} \gamma_{t-1} \\ \delta_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ u_t \end{bmatrix}$$

which is in the familiar form (notice that the covariance matrix of the redefined error is singular). Typical sample paths for structure (93) are shown in figure 9.2.

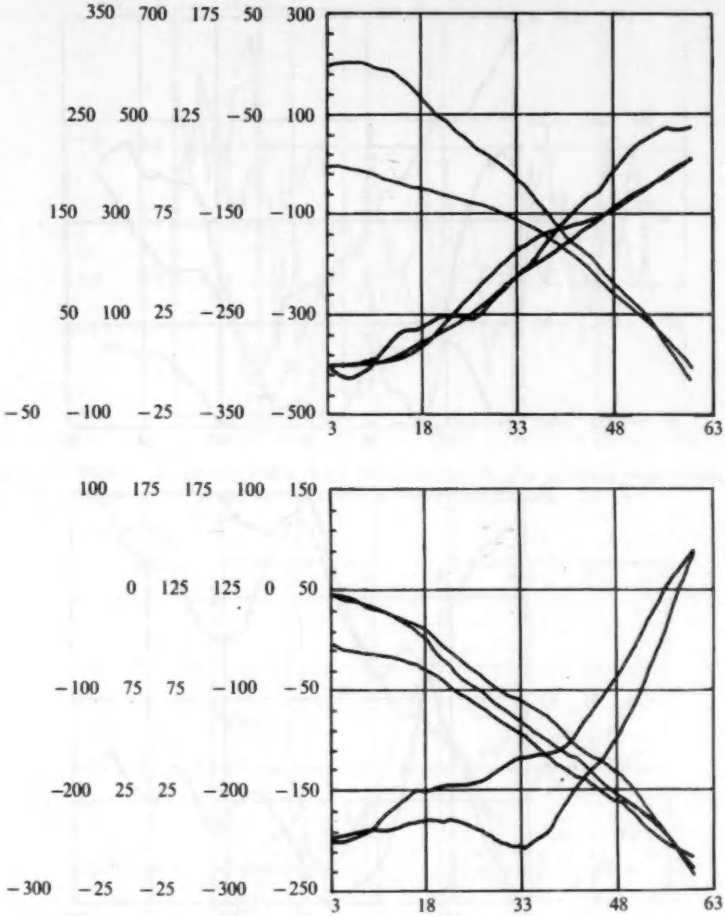

Figure 9.2 Time Series Sample Paths of the Structure $\beta_t = 2\beta_{t-1} - \beta_{t-2} + u_t$. The initial conditions are zero for all paths; u_t is normal white disturbances.

It is apparent from the figure that the *a priori* assumption about the variation of β_t is much different than the one used in posing structure (93).

It is obvious that many different ARIMA structures can be imposed on β_t. Figure 9.3 shows sample paths from a (0, 1, 2) ARIMA process of the form

(97) $$\beta_t = \beta_{t-1} + u_t$$

(98) $$u_t - 1.2u_{t-1} + 0.32u_{t-2} = \varepsilon_t.$$

Figure 9.4 shows sample paths from a (1, 1, 0) ARIMA process and figure 9.5 shows sample paths from a (0, 1, 1) ARIMA process.

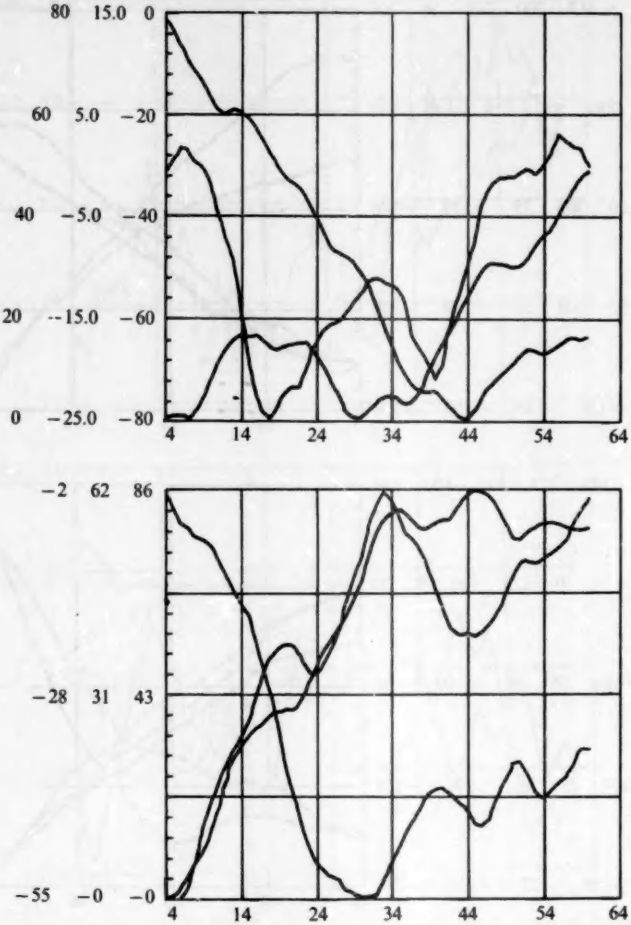

Figure 9.3 Time Series Sample Paths of the Structure $\beta_t = \beta_{t-1} + u_t$ with $u_t - 1.2u_{t-1} + 0.32u_{t-2} = \varepsilon_t$. The initial conditions are zero for all paths; ε_t is normal white disturbances.

The difficulty of identifying the transition matrix T, or in general the *a priori* structure of the time variation now becomes clear. For different ARIMA models different structure is posed *a priori* on the parameters. Even if the analyst feels strongly about one particular kind of variation, there are probably more than one ARIMA models that give sample paths with the desired character. The dilemma to the analyst is not an easy one to resolve. It is similar to the problem of isolating the kinds of variables to be included in an econometric model.

Notice that once the structure is imposed then the theory developed earlier in the paper can be used to estimate the time series β. The following procedure is suggested to resolve the identifiability problem.

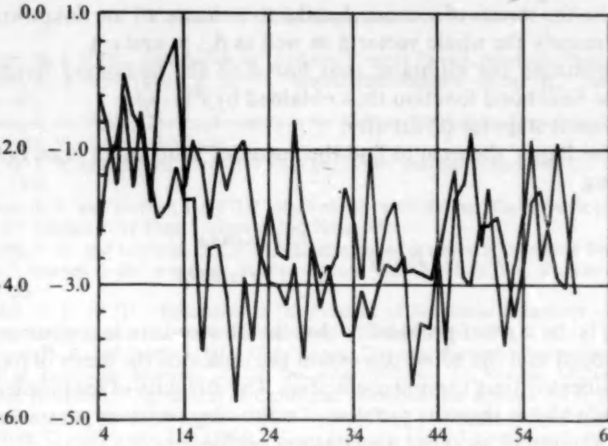

Figure 9.4 Time Series Sample Paths of the Structure $\beta_t = \beta_{t-1} + u_t$ with $u_t = \varepsilon_t - 0.6\varepsilon_{t-1}$. The initial conditions are zero for all paths; ε_t is normal white disturbances.

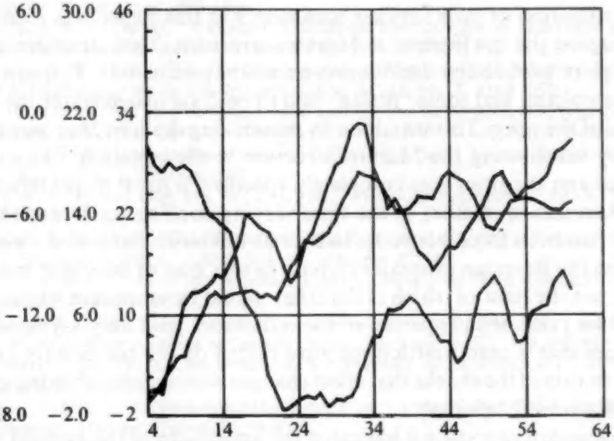

Figure 9.5 Time Series Sample Paths of the Structure $\beta_t = \beta_{t-1} + u_t$ with $u_t = 0.6u_{t-1} + \varepsilon_t$. The initial conditions are zero for all paths; ε_t is normal white disturbances.

Procedure

(a) By careful thinking about the problem isolate a finite set of structures of the form (92) or (93) or any other appropriate character. Denote by $s_i (i = 1, \ldots, m)$ the ith chosen structure.

(b) Assign, *a priori*, a probability mass function on the set $S = \{s_i\}$.

(c) For each structure s_i the likelihood function of the data is $l(y; s_i)$, and will be a function of the unknown as yet vector β and the other unknown parameters of s_i (e.g. the β_0, σ_ε and σ_u of section 4).

(d) Use the results of sections 4 and 5 to estimate all the unknown quantities of $l(y; s_i)$ (namely the whole vector β as well as β_0, σ_ε and σ_u).

(e) Substitute the estimates thus found to the likelihood function $l(y; s_i)$. Denote the likelihood function thus obtained by $l^0(y; s_i)$.

(f) Repeat steps (c)–(e) for all i.

(g) Use Bayes' theorem to find the posterior probability mass function on S as following

$$p(s_i|y) = \frac{l^0(y; s_i)p(s_i)}{\Sigma_{i=1}^m l^0(y; s_i)p(s_i)} \tag{99}$$

where $p(s_i)$ is the *a priori* probability that the ith structure is the correct one.

It is hoped that the above discussion has indicated the limits of methods that attempt to identify time varying coefficients. The difficulty of the problem is orders of magnitude higher than the problem of estimating constant parameters because there is an infinity of ways the nonconstant coefficients can vary.

10. Summary and Conclusions

The estimation of time varying parameters in this paper was made possible by assigning to the coefficients a Markov structure. This structure essentially imposed a prior probability distribution on all the parameters. This prior was not specified completely and some "fitting" had to be done to determine the unknown parameters of the prior. This was done by minimizing the sum of squared residuals obtained by substituting the Markov structure in the equation. Once the fitting was finished and the prior was completely specified, a prior to posterior analysis yield the Bayesian estimators of the time varying coefficients. It was shown that if the prior had been fitted perfectly, so that its unknown parameters were known exactly, then the Bayesian estimators would have a host of desirable small sample properties, not the least of which is that they would be minimum variance. Given the inaccurate prior, the properties of the estimators hold only asymptotically. It was also seen that if exact restrictions were placed on the coefficients, so that the covariance matrix of the shocks that effect changes was singular, then the estimators so obtained were still optimal.

The discussion of section 9 indicated the limitations of the method presented here, as well as of any method that attempts to identify time varying coefficients. The Bayesian approach was seen to be a possible answer to the dilemma of which prior structure to use. The final choice of method and structure rests on the analyst and should be dictated by the goals of his analysis.

All the results derived in this paper hold when the matrix X does not include lagged values of the endogenous variable. Research must still be done on estimation methods that take this fact into account. Furthermore, much experimental work is needed to obtain information about the small sample properties of time varying parameter estimators.

National Bureau of Economic Research

BIBLIOGRAPHY

[1] Åström, K. J. (1970), *Introduction to Stochastic Control Theory*, Academic Press, 1970.

[2] Åström, K. J. and Eykhoff, P. (1971), "System Identification—A Survey," *Automatica*, Vol. 7, May 1971.

[3] Belsley, D. A. (1973), "The Applicability of the Kalman Filter in the Determination of Systematic Parameter Variation," this issue.

[4] Box, G. E. P. and Jenkins, G. M. (1970), *Time Series Analysis, Forecasting and Control*, Holden-Day, 1970.

[5] Brown, R. L. and Durbin, J. (1971), "Methods of Investigating Whether a Regression Relationship is Constant Over Time," Unpublished Note, 1971.

[6] Burnett, J. D. and Guthrie, D. (1970), "Estimation of Stationary Stochastic Regression Parameters," *Journal of the American Statistical Association* (*JASA*), Vol. 65, December 1970, pp. 1547–53.

[7] Cooley, T. F. (1971), "Estimation in the Presence of Sequential Parameter Variation," Ph.D. Thesis, Department of Economics, University of Pennsylvania, 1971.

[8] Duncan, D. B. and Horn, S. D. (1972), "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," *JASA*, December 1972, Vol. 67, Number 340, pp. 815–21.

[9] Heffes, H. (1966), "The Effects of Erroneous Models on the Kalman Filter Response," *IEEE Transactions on Automatic Control* (Short Papers), Vol. AC-11, pp. 541–43, July 1966.

[10] Hildreth, C. and Houck, J. P. (1968), "Some Estimators for a Linear Model with Random Coefficients," *JASA*, Vol. 63, June 1968, pp. 584–95.

[11] IEEE (1971), Special Issue on Linear Quadratic Gaussian Problem, *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 6, December 1971.

[12] Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, Clarendon Press, 1961 and 1966.

[13] Johnson, R. A. (1967), "An Asymptotic Expansion for Posterior Distributions," *Annals of Math. Statistics*, 38, 1967, pp. 1899–1906.

[14] Kashyap, R. L. (1970), "Maximum Likelihood Identification of Stochastic Linear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-15, No. 1, February 1970.

[15] Mehra, R. K. (1970), "On the Identification of Variances and Adaptive Kalman Filtering," *IEEE Transactions on Automatic Control*, Vol. AC-15, No. 2, April 1970.

[16] ———. (1971), "On Line Identification of Linear Dynamic Systems with Applications to Kalman Filtering," *IEEE Transactions on Automatic Control*, Vol. AC-16, No. 1, February 1971.

[17] ———. (1972), "Approaches to Adaptive Filtering," *IEEE Transactions on Automatic Control*, Vol. AC-17, No. 5, October 1972.

[18] McGee, V. E. and Carleton, W. T. (1970), "Piecewise Regression," *JASA*, Vol. 65, No. 331, September, 1970, pp. 1109–24.

[19] Quandt, R. E. (1958), "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *JASA*, Vol. 53, 1958, pp. 873–80.

[20] ———. (1960), "Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes," *JASA*, Vol. 55, 1960, pp. 324–30.

[21] Rao, C. R. (1965), "The Theory of Least Squares when the Parameters are Stochastic and its Application to the Analysis of Growth Curves," *Biometrica*, 52, 1965, pp. 447–58.

[22] Rosenberg, B. (1967), "Regression in the Presence of Stochastically Varying Parameters," Unpublished mimeo.

[23] ———. (1968a), "Varying-Parameter Estimation," Unpublished Ph.D. Thesis, Department of Economics, Harvard University, 1968.

[24] ———. (1968b), "An Atomistic Approach to Linear Regression," Unpublished mimeo.

[25] ———. (1972), "The Estimation of Stationary Stochastic Regression Parameters Reexamined," *JASA*, Vol. 67, No. 339, September 1972, pp. 650–54.

[26] Sage, A. P. and Melsa, J. L. (1971), *Estimation Theory with Applications to Communications and Control*, McGraw-Hill, 1971.

[27] Swamy, P. A. V. B. (1970), "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, No. 2, March 1970, pp. 311–23.

[28] Theil, H. (1971), *Principles of Econometrics*, John Wiley, 1971.

[29] Wieslander, J. and Wittenmark, B. (1971), "An Approach to Adaptive Control Using Real Time Identification," *Automatica*, Vol. 7, May 1971.

[30] Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley, 1971.