THE ESTIMATION OF PARAMETRIC CHANGE IN TIME-SERIES MODELS

by

J. M. Kaldor

A thesis submitted to the
Australian National University
for the degree of Master of Arts
August, 1978

STATEMENT

The contents of this thesis are entirely my own work, except where otherwise indicated.

John Kald

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The work for this thesis was undertaken at the Centre for Resource and Environmental Studies, Australian National University, with the financial support of a Commonwealth Postgraduate Research Award. I thank my supervisors, Professor P.C. Young of the Centre, Dr. D.F. Nicholls of the Department of Statistics, S.G.S., and Professor E.J. Hannan of the Department of Statistics, R.S.S.S., Australian National University, for their advice and assistance. I also thank a number of members of the Centre for their support and encouragement during the preparation of the thesis. Finally thanks are due to the typist, Mrs. M. Mitchell.

ABSTRACT

This thesis examines methods for detecting structural change in parametric time-series models. This detection is accomplished through the use of random walk models of the parameter variation. Although the model of main interest is the transfer function model, the methods developed are largely adaptations of procedures used for regression models, as the exact theory for the time-series case is generally too complex. An instrumental variable smoothing algorithm for estimating parametric change is developed, and is shown to provide good estimates of the variation. Other aspects of the procedure are also discussed,including the estimation of the statistics of the parameter variation. Finally, some computer simulations and analyses of real data are provided. These illustrate some of the main points discussed in the thesis.

CHAPTER 1 : INTRODUCTION

1.1 The Model

In the following, we will be examining the fitting of models to data where it is both meaningful and useful to consider the measured data as being generated by a parametric model, whose parameters may be either constant over the observation period or functions of time. Suppose initially that the parameters are constant. Then it is hoped that, for a sample of size N, a model of the form

$$y_k = f(u_k^{(1)},\ldots,u_{k-n}^{(1)};\ldots,u_k^{(m)},\ldots,u_{k-n}^{(m)};\underset{\sim}{\theta}) + e_k,\ k = 1,2,\ldots,N \quad (1.1.1)$$

for some f, n, will explain the relationship between the m scalar *inputs*, $u_k^{(i)}$, i = 1,2,...,m; and the scalar *output* y_k. The p-dimensional *parameter* vector $\underset{\sim}{\theta}$ is unknown, and it is the estimation of the elements of this vector which concerns us.

In equation (1.1.1) e_k is a random quantity which cannot be measured. It is assumed to represent the lumped effect of measurement error and other stochastic disturbances interfering with the exact establishment of the *noise-free output* $x_k \overset{\Delta}{=} f(.)$. Although the question of the form which f should take in various situations is of great importance, it will not be discussed here : it is assumed that f has a known form.

To be more specific, two basic models will be considered. They could be generalized into the one model. However for the purposes of this exposition, they will be examined separately.

(I) The regression model

In (1.1.1), let

$$f(.) = \underset{\sim}{u}_k^T \underset{\sim}{\theta}, \text{ where } \underset{\sim}{u}_k^T = (u_k^{(1)}, \ldots, u_k^{(m)})$$

θ is then an m-vector of unknown parameters : in this case $p = m$.

(II) The transfer function model

In (1.1.1), let

$$f(.) = \frac{B(z^{-1})}{A(z^{-1})} u_k$$

Here it is assumed that there is only one input, u_k.

$$A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n}$$

$$B(z^{-1}) = b_0 + b_1 z^{-1} + \ldots + b_n z^{-n},$$

where z^{-1} is the backward shift operator $z^{-1}x_k = x_{k-1}$, defined on all functions of the integers; $z^{-i} = (z^{-1})^{i}$.

$\underset{\sim}{\theta}^T \overset{\Delta}{=} (a_1,\ldots,a_n,b_0,\ldots,b_n)$ is the vector of unknown parameters. In this case, $p = 2n+1$. Where the meaning is unambiguous these polynomials in z^{-1} may be abbreviated to A,B.

Although the transfer function $\frac{B}{A}$ is strictly only defined as a quotient in the Laplace domain, where $z^{-1} = e^{-st}$, t being the sampling interval, we can consider x_k as the solution to the difference equation $A(z^{-1})x_k = B(z^{-1})u_k$. This eliminates any possible objections to the usage of $\frac{B}{A}$.

In both models, the e_k are independent and indentically distributed (i.i.d.) random variables, with mean zero and variance σ^2; they are uncorrelated with the inputs $u_k^{(i)}$, $1 = 1,2,\ldots,m$.

A number of generalizations can be made to both of the models. For example, correlation amongst the e_k's could be introduced; the number of inputs u_k in model II, or the number of outputs y_k in both models could be increased; or the explanatory variables in model I could be assumed to be measured with error. It will be indicated subsequently how some of the modifications affect what is to follow. However at this stage, the simpler models set out above will suffice.

1.2 Parameter Variation

The estimation of the unknown parameters in both the models I and II has been dealt with extensively in the statistical and other literature. Kendall and Stuart (1961), for model I, and Box and Jenkins (1970) and Young (1974; 1976) for model II, are some of many references. However, most of the work in the area has been carried out under the assumption that the parameters remain constant over the observation interval. Often, in situations where the measurements are made at successive points in time, it is reasonable to suppose that the relationships *do* change over time; or at least it would be of interest to ascertain *if* they do, particularly where there is some *a priori* reason to believe this to be the case. As a result, it is useful to generalize the two models given, by replacing $\underset{\sim}{\theta}$ by $\underset{\sim}{\theta}_k$, and A, B by A_k, B_k in model II. We are still interested in estimating $\underset{\sim}{\theta}_k$, $k = 1,2,\ldots,N$, but the problem has now become more complex. The number of unknown parameters is now $pN + 1$, which is a monotonically increasing function of the sample size N. From the point of view of statistical analysis, such a situation is unsatisfactory, since there are more parameters to estimate than there are observations.

Of course it must be emphasised that the problem may not be as complicated as this in practice. For example, if the estimation procedure used indicated some specific pattern of parameter variation with time, this variation could be related to some other variable. The relationship would then be built into the model, eliminating the need for a different parameter at each time point, possibly using a similar approach to the intervention analysis of Box and Tiao (1975). Other simplifications may occur, as in Young (1969), where the variation of a highly-time-varying parameter is largely accounted for by modelling the parameter $\underset{\sim}{\theta}_k$ as $\underset{\sim}{\theta}_k = T_k\underset{\sim}{\theta}^*_k$, where T_k is a matrix of highly varying, but measurable, state variables, and $\underset{\sim}{\theta}^*_k$ is a very slowly, and hence more easily modelled, time varying parameter.

In general, however, there will be a problem of estimating time varying parameters, and we now consider a number of ways of approaching this problem. Before doing so it is useful to distinguish between off-line (or block) and on-line (or recursive) procedures. An on-line procedure is one where the estimate of a parameter at a given point in time can be obtained directly from the current data, and the estimate at the previous point in time. Block procedures are those where all data must be processed at each time point to obtain the estimate at that time point. We will now discuss a number of estimation methods for time varying parameters.

1.2.1 Non-uniform data weighting

This procedure has been used in engineering applications (Young, 1969; Jazwinski, 1970). In an off-line estimation for a parameter vector assumed constant, all data carry equal weight with respect to the estimation, in the above two models. If, however, we wanted to assume that the parameter may be different at each time point, we can, at the expense of estimation error variance, consider 'current' data as carrying more weight in the estimation, in some way.

There are various ways in which this can be accomplished. The simplest is to estimate the parameters at a given time only using the data in a certain interval about that time. Thus, the estimate at time k would be obtained only from data in the time interval (k-t, k+t), where $k-t \geq 1$, and $k+t \leq N$. A more sophisticated alternative is to exponentially weight past data, so that they carry less weight as they become 'older'. The main difficulty with this type of scheme is the arbitrary nature of the weighting which will, of necessity, result in general. Furthermore, a stationary weighting procedure, that is, one which weights in the same pattern about each time point, may be too restrictive to detect all types of parameter variation. On the other hand, it is difficult to develop any non-stationary procedure.

1.2.2. Stationary stochastic parameters

This approach has been considered quite extensively in the econometric literature (Hildreth and Houck, 1968; Swamy, 1971; Rosenberg, 1972; Pagan, 1978). It has been applied mostly in econometric models, which are, generally, multivariate regressions with some explanatory variables (inputs) measured with error. The procedure is to suppose that the values of the unknown parameter $\underset{\sim}{\theta}_k$, $k = 1,\ldots,N$ are a realization of a stochastic process $\underset{\sim}{\theta}_k = \underset{\sim}{\theta} + \xi_k$, where ξ_k is a mean-zero, wide sense stationary stochastic process. The earlier work (Swamy; Hildreth and Houck) took the ξ_k as i.i.d., while more recently, they have been modelled as an autoregression (Pagan). Although this allows time dependence, it still implies that the parameters are estimated with an identical distribution at each time point, so that large deviations may not be detected very clearly. Our aim here is to employ some methods where such detection is accomplished.

1.2.3 Non-stationary stochastic parameters

This is the method which we shall be concerned with in the remainder of this thesis. No rigorous attempt will be made at this stage to *define* the type of variation we could hope to model in this way. However the following general assumptions (based on Bennett, 1976) will prove helpful.

(i) The parameter variation follows some sort of 'pattern' which is not totally random, whether stochastic or deterministic. Thus the parameters are not a realization of a white noise process.

(ii) The parameter variation is independent of the observation error e_k, in the two models I and II.

Taking these assumptions into consideration, it is appropriate to choose a stochastic process which is not too restrictive. Here once again, it is difficult to be rigorous. Nevertheless what is meant, roughly, is that conceivable parameter variation (i.e. sample paths) does not lie too far into the tails of the distribution of the stochastic process. At the same time, the process should have some memory, so that past data is not altogether discarded. The first requirement leads us to a non-stationary process, while the second, combined with the need for simplicity , suggests the use of a Markov process. The class of processes we choose are the random walks : Markov processes with state-space p-dimensional Euclidean space, and variance unboundedly increasing with time.

The major aim of this thesis is to consider the estimation of time variable parameter (non-stationary) dynamic systems in which the parameter variation can be described by a random walk of some kind. In Chapter 2, the random walk model will be examined in more detail, and various approaches to the estimation of the parameters will be considered. In Chapters 3 and 4, algorithms will be derived for estimating the parameters as a realization of a random walk in the models I and II, respectively. A number of additional details concerning the utilization of the algorithms will be discussed in Chapter 5. In Chapter 6, the results of some simulations and analyses of real data are reported, and Chapter 7 mentions some extensions to the procedures discussed in the thesis, and outlines some possible future work that could be carried out.

It may be noted that throughout the following chapters, there is a dichotomy in the approach being taken. In places, it will appear that the aim of the methodology being developed is to track any parametric variation which may occur. Elsewhere, a more rigorous statistical approach will be taken, and the underlying parametric variation will be assumed to be a random walk. Of course, it may be said that *any* statistical modelling implicitly involves such a dichotomy. However it is preferred here to make it explicit.

Young (1969; 1974), Norton (1975) and Garbade (1977) are the most important sources for, and are most closely related to, this thesis.

CHAPTER 2 : THE RANDOM WALK MODEL

2.1 Background

The use of the random walk in the context of varying parameter models appears to have been first suggested by Kopp and Orford (1963), who used it to track parameters in an adaptive control system using a re-linearized or extended Kalman Filter. Lee (1964) applied a random walk to obtain estimates of parameters varying in time, and Young (1965; 1969) expanded on this in an instrumental variable context. As mentioned in Section 1.2.2, autoregressive type schemes have appeared in the econometric literature. However, in this area Garbade (1977) seems to be the first to track variation of regression parameters, rather than model them in a stationary manner. Norton (1975) introduced the added advantage of smoothing (see Section 2.3) in a set-up similar to model II.

2.2 Types of Random Walk

The *simple random walk* (RW) has appeared in the context of Section 2.1 most frequently (Lee, 1964; Young, 1965; 1969; Bennett, 1976; Norton, 1975; Garbade, 1977). Here we take as the model of parameter variation

$$\underset{\sim}{\theta}_k = \underset{\sim}{\theta}_{k-1} + \underset{\sim}{\nu}_k$$

where $\{\underset{\sim}{\nu}_k\}_{k=1}^{N}$ is an i.i.d. sequence of random vectors with mean zero, and variance-covariance matrix Q. This model has the advantage of simplicity, both in concept and implementation. However, the model has a definite restriction in situations where large changes may occur over small time intervals. The value of Q required to track such changes may mean that the random walk is very 'jagged' (See Section 6.1). To overcome this difficulty, Norton (1976) has employed the *integrated random walk* (IRW). Here it is supposed that the first difference of the process is a simple random walk. It is necessary to augment the parameter vector $\underset{\sim}{\theta}_k$ by the increment vector $\underset{\sim}{S}_k$, so that the number of parameters is doubled. The model is now

$$\begin{pmatrix} \underset{\sim}{\theta}_k \\ \\ \underset{\sim}{S}_k \end{pmatrix} = \Phi \begin{pmatrix} \underset{\sim}{\theta}_{k-1} \\ \\ \underset{\sim}{S}_{k-1} \end{pmatrix} + \Gamma \underset{\sim}{\nu}_k \qquad (2.2.1)$$

where

$$\Phi = \begin{pmatrix} I_p & I_p \\ \\ 0 & I_p \end{pmatrix} \quad \Gamma = \begin{pmatrix} 0 \\ \\ I_p \end{pmatrix}$$

$\underset{\sim}{\nu}_k$ is as above, and I_p is the $p \times p$ identity matrix.

Clearly, it would also be possible to use random walks where the second or even higher difference was a random walk, with a corresponding increase in the size of the parameter vector.

Something of a compromise between the IRW and the RW is the *smoothed random walk* (SRW) (Young and Kaldor, 1978). Here the effect of the random walk increments occurring in the IRW is somewhat diminished by the inclusion of the coefficients α_i, which are typically in the range 0.9 - 1.0. Then the model of parameters variation is as in (2.2.1), with now

$$\Phi = \begin{pmatrix} \alpha & \beta \\ & \\ 0 & I_p \end{pmatrix}$$

where

$$\alpha = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_p)$$

$$\beta = \text{diag}(1-\alpha_1, 1-\alpha_2, \ldots, 1-\alpha_p)$$

The three types of random walk all have zero mean, and if we assume $\underset{\sim}{\theta}_0 \equiv \underset{\sim}{0}$, $S_0 \equiv \underset{\sim}{0}$, the random walks have variances[†], respectively

$$V(\underset{\sim}{\theta}_k{}^{(RW)}) = k^2 Q$$

$$V(\underset{\sim}{\theta}_k{}^{(IRW)}) = (\sum_{i=1}^{k-1} i^2)Q$$

$$V(\underset{\sim}{\theta}_k{}^{(SRW)} = (\alpha^{k-1} + \sum_{i=3}^{k} {}^{k-i}\beta)Q(\alpha^{k-1} + \sum_{i=1}^{k} \alpha^{k-i}\beta)$$

$$+ \sum_{j=3}^{k} (\sum_{i=j}^{k} \alpha^{k-i}\beta)Q(\sum_{i=j}^{k} \alpha^{k-i}\beta)$$

[†] Here, and subsequently, we may use the word 'variance' to denote the variance-covariance matrix of a vector random variable, if the meaning is unambiguous.

The use of any particular one of these models should depend on the context. In general, because of the 'parameter tracking' approach being taken here, a selection of these models can be employed, and further investigations carried out in accord with the results. This point is considered in more detail in Chapter 6.

2.3 Parameter Estimation in a Random Walk Model

As a result of the discussion in Section 1.1 and 2.2, the model we now consider is an *observation equation*

$$y_k = x_k + e_k \qquad (2.3.1)$$

where all quantities are defined as in Section 1.1, with $\underset{\sim}{\theta}$ replaced by $\underset{\sim}{\theta}_k$ in (1.1.1); and a *parameter evolution equation*

$$\underset{\sim}{\theta}_k = \Phi\underset{\sim}{\theta}_{k-1} + \Gamma\underset{\sim}{\nu}_k \qquad (2.3.2)$$

where all quantities are defined as in Section 2.2, $\underset{\sim}{\theta}_k$ being augmented to include $\underset{\sim}{S}_k$ in the IRW and SRW models, and Φ, Γ depending on the random walk chosen.

There are a number of different approaches that can be taken to estimate $\underset{\sim}{\theta}_k$. Since it has been postulated that the parameters are random variables, the most complete knowledge one can have of them is their exact density function, if we assume distributions absolutely continuous with respect to Lebesgue measure throughout. This requires knowledge of the density functions of $\underset{\sim}{\theta}_0$, and $\underset{\sim}{\nu}_k$, e_k for $k = 1,2,\ldots,N$. We will denote densities by p(.) where the argument is the random variable whose density is being represented. Now the Chapman-Kolmogorov equation (Jazwinski, 1970) gives

$$p(\underset{\sim}{\theta}_k) = \int p(\underset{\sim}{\theta}_k|\underset{\sim}{\theta}_{k-1})p(\underset{\sim}{\theta}_{k-1})d\underset{\sim}{\theta}_{k-1}$$

as the equation of evolution with time of the densities $p(\underset{\sim}{\theta}_k)$, $k = 1,2,\ldots,N$. While the process $\{\underset{\sim}{\theta}_k\}_{k=1}^N$ is not observed, a related process $\{Y_k\}_{k=1}^N$ is observed. Y_k is the random variable whose realization is denoted by y_k in (2.3.1). Thus, without additional *a priori* information, the best that can be done is to obtain information about $p(\underset{\sim}{\theta}_k)$ from some subset of $\{y_1,y_2,\ldots,y_N\}$.

It is clear at this stage that the problem is cast in exactly the same framework as the discrete-time state estimation problem (Kalman, 1960). The latter situation is concerned with estimating the value (state) of a discrete time stochastic process $\{\underset{\sim}{x}_k\}_{k=1}^N$. The major difference between the two problems arises from the fact that the states have physical meaning, and the stochastic process describing their evolution is usually derived from physical principles. In the present situation, however, the parameter evolution is described by the random walk, which it is hoped will accommodate the true behaviour of the parameter, even though a 'typical' realization of the random walk may not resemble the parameter variation at all.

Because of this, the choice of the subset of $\{y_1,y_2,\ldots,y_N\}$

to be used in the estimation of $\underset{\sim}{\theta}_k$, is constrained. For example, in a state estimation problem it may be possible to make some inference about $p(\underset{\sim}{x}_\ell)$ on the basis of $\{y_1,y_2,\ldots,y_k\}$ where $k < \ell$ (this is the *prediction* problem considered by Kalman, 1960) if physical knowledge of the $\underset{\sim}{x}_k$ process provides information on $p(\underset{\sim}{x}_{k+1}),\ldots,p(\underset{\sim}{x}_\ell)$. In the parameter estimation situation, however, the use of the random walk means that the most that can be known about $p(\underset{\sim}{\theta}_\ell)$ on the basis of $\{y_1,y_2,\ldots,y_k\}$ is contained in $p(\underset{\sim}{\theta}_k)$.

Therefore, we will always restrict attention to the problem of making inferences about $p(\underset{\sim}{\theta}_\ell)$ on the basis of $\{y_1,y_2,\ldots,y_k\}$, where $k > \ell$. It should be noted that if $k = \ell$, this corresponds to the *filtering* problem of state estimation. If $k > \ell$ it corresponds to the *smoothing* problem (Kalman, 1960).

Taking a Bayesian approach because only one realization of the observation process $\{Y_k\}_{k=1}^N$ is available, the density function of interest is now $p(\underset{\sim}{\theta}_\ell|\underset{\sim}{Y}_k)$ where now we define $\underset{\sim}{Y}_k^T = (Y_1,Y_2,\ldots,Y_k)$. This gives all obtainable information concerning the density conditional on the observed data, and constitutes the complete (Bayesian) solution to the problem (Cox, 1964). This density is called the *a posteriori* density, and is given by Bayes theorem as

$$p(\underset{\sim}{\theta}_\ell|\underset{\sim}{Y}_k) = \frac{p(\underset{\sim}{Y}_k|\underset{\sim}{\theta}_\ell)\ p(\underset{\sim}{\theta}_\ell)}{p(\underset{\sim}{Y}_k)}$$

It now remains to be decided what will be called an estimate of $\underset{\sim}{\theta}_{\ell}$, if $p(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$ is known. The choice can be made by minimizing the expected value of some loss function L.[†] It can be shown (Sherman, 1955; quoted in Cox, 1964), that if $p(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$ is symmetric about its mean, and unimodal, then E(L) is minimized by taking as an estimate the *conditional mean* $E(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$. On the other hand, Sage and Melsa (1971) show that, for a quadratic loss function, the expected loss is minimized by the conditional mean, and for a loss function uniform in a symmetric interval about the origin and zero elsewhere, the expected loss is minimized as the interval size → 0 by $\underset{\sim}{\theta}_{\ell}$ such that $p(\underset{\sim}{\hat{\theta}}_{\ell}|\underset{\sim}{Y}_{k}) = \max\limits_{\underset{\sim}{\theta}_{\ell}} p(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$ - the *maximum a posteriori* estimate. Then if $p(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$ is unimodal and symmetric, these two estimates will coincide. In particular, this occurs when $p(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$ is Gaussian.

One of the main difficulties in evaluating any solution to the problem of parameter tracking lies in the choice of estimation criteria. If the parameters were actually varying as a random walk, and the density $p(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k})$ could be obtained exactly, then the conditional mean estimate $\underset{\sim}{\hat{\theta}}_{\ell}$ will be unbiased, since

$$E(\underset{\sim}{\hat{\theta}}_{\ell}-\underset{\sim}{\theta}_{\ell}) = E(E(\underset{\sim}{\theta}_{\ell}|\underset{\sim}{Y}_{k}) - \underset{\sim}{\theta}_{\ell}) = E(\underset{\sim}{\theta}_{\ell}) - E(\underset{\sim}{\theta}_{\ell}) = 0$$

It will also be minimum variance, since it minimizes the quadratic loss function. Pagan (1978) considers the likelihood obtained

[†] That is, some function L of the difference between the estimate and the true value of the parameter, such that L(0) = 0, and for convex p, $p(\alpha)\geq p(\beta)\geq 0$ implies $L(\alpha)\geq L(\beta)\geq 0$.

under Gaussian assumptions on $\underset{\sim}{\nu}_k$, e_k and $\underset{\sim}{\theta}_0$, with parameters following a stationary autoregression, in model I. He asserts that the maximum likelihood estimates of $E(\underset{\sim}{\theta}_0)$, $V(\underset{\sim}{\theta}_0)$, Q, σ^2 and Φ are consistent and obey a central limit theorem. However his proof does not include the non-stationary random walk considered here. Moreover, since the true parameter variation is not generally assumed known, these properties are not necessarily useful. It may be that the best criteria available in general is a sum of squared or absolute deviations. For simulated data, these deviations can be the difference between estimates and known values of time-varying parameters. For real data, they can be j-step ahead prediction errors.

Because parameter estimation is usually done off-line, we have the opportunity to make use of as much data as possible. Thus the estimate $\hat{\underset{\sim}{\theta}}_\ell$, based on $p(\underset{\sim}{\theta}_\ell|\underset{\sim}{Y}_N)$, should be used wherever possible. In the case where we have a quadratic loss function, and an estimate $\underset{\sim}{\theta}_\ell$ is required which is linear in $y_1,y_2,\ldots,y_N$, the advantage of this can be seen very clearly using the 'innovations approach' (Kailath and Frost, 1968; Aasnaes and Kailath, 1973).

As before, we take $\underset{\sim}{Y}_\ell^T = (Y_1,Y_2,\ldots,Y_\ell)$, for $\ell = 1,2,\ldots,N$. $\underset{\sim}{Y}_\ell$ can be orthogonalized by defining the *linear innovations*

$$\varepsilon_k = Y_k - \hat{Y}_{k|k-1},\ k=1,2,\ldots,N,\ \text{where } Y_{1|0} = 0 \qquad (2.3.3)$$

$\hat{Y}_{k|k-1}$ denotes the minimum quadratic loss, linear estimate of Y_k based on $\{y_1,y_2,\ldots,y_{k-1}\}$. If we define an inner product on Ω_k, the linear space spanned by $\{Y_1,Y_2,\ldots,Y_k\}$, by EXY, then

$\hat{\underset{\sim}{\theta}}_{\ell|k}$ can be seen as the orthogonal projection of $\underset{\sim}{\theta}_\ell$ onto Ω_k.

It is given by

$$\hat{\underset{\sim}{\theta}}_{\ell|k} = \sum_{t=1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\} \ \varepsilon_t \tag{2.3.4}$$

(suppose $k \geqslant \ell$).

Let

$$R(\ell) = E(\underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|k})(\underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|k})^T,$$

the variance of the smoothed estimate;

$$S(\ell) = E(\underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|\ell})(\underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|\ell})^T,$$

the variance of the filtered estimate.

Then

$$\hat{\underset{\sim}{\theta}}_{\ell|k} = \hat{\underset{\sim}{\theta}}_{\ell|\ell} + \sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\} \ \varepsilon_t$$

so that

$$\underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|\ell} = \underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|k} + \sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\} \ \varepsilon_t$$

Multiplying each side by its transpose, and taking expectation gives

$$\begin{aligned} S(\ell) = R(\ell) &+ E\ (\underset{\sim}{\theta}_\ell - \hat{\underset{\sim}{\theta}}_{\ell|k})(\sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\}\varepsilon_t)^T \\ &+ E(\sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\}\varepsilon_t)(\sum_{t=\ell+1}^{k} E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\}\varepsilon_t)^T \\ + R(\ell) + 0 &+ E(\sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\} \ \varepsilon_t)(\sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t{}^2)\}\varepsilon_t)^T \end{aligned}$$

since the error in the orthogonal projection onto Ω_k is orthogonal to Ω_k. Furthermore

$$S(\ell) = R(\ell) + \sum_{t=\ell+1}^{k} \{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t^2)\}\{E(\underset{\sim}{\theta}_\ell \varepsilon_t)/E(\varepsilon_t^2)\}^T E(\varepsilon_t^2)$$

from the orthogonality of $\varepsilon_{\ell+1}, \varepsilon_{\ell+2},\ldots,\varepsilon_k$.

Thus

$$S(\ell) = R(\ell) + \sum_{t=\ell+1}^{k} E(\underset{\sim}{\theta}_\ell \varepsilon_t)E(\underset{\sim}{\theta}_\ell \varepsilon_t)^T/E(\varepsilon_t^2)$$

Therefore $S(\ell) > R(\ell)$, since the second term is a symmetric, positive definite matrix. Thus the filtering error variance is at least equalled, and normally decreased, by smoothing.

CHAPTER 3 : ALGORITHMS FOR ESTIMATION IN MODEL I

3.1 Introduction

In this chapter we will derive a number of algorithms for the computation required in the estimation procedures discussed in the previous chapter. Some other algorithms will also be discussed.

Referring to equation (2.3.1), (2.3.2), we have

$$y_k = \underset{\sim}{u}_k^T \underset{\sim}{\theta}_k + e_k \tag{3.1.1}$$

$$\underset{\sim}{\theta}_k = \Phi\underset{\sim}{\theta}_{k-1} + \Gamma\underset{\sim}{\nu}_k \tag{3.1.2}$$

It is possible to distinguish two sets of (possibly overlapping) conditions which will result in the same estimation procedures in model I.

1) The densities $p(\underset{\sim}{\theta}_o)$, $p(e_k)$, $p(\underset{\sim}{\nu}_k)$, $k = 1,2,\ldots,N$ are all Gaussian, and the estimate required is either the maximum *a posteriori* or the conditional mean estimate. Then, because $\underset{\sim}{\theta}_\ell$ is linearly related to $\underset{\sim}{\theta}_o$, and e_k, $\underset{\sim}{\nu}_k$, $k = 1,2,\ldots,N$,the conditional density $p(\underset{\sim}{\theta}_\ell|Y_k)$ will also be Gaussian. It is thus completely characterized by its mean and variance, and the mean also gives the maximum of the density. The conditional mean, which is to be used as an estimate, will be linear in $y_1,y_2,\ldots,y_k$, so that a generalisation of the condition is to suppose that the conditional expectation $E(\underset{\sim}{\theta}_\ell|\underset{\sim}{Y}_k)$ is

linear in $y_1, y_2, \ldots, y_k$.

2) The densities are not necessarily Gaussian; a linear function of $y_1, y_2, \ldots, y_k$ is required to estimate $\hat{\underset{\sim}{\theta}}_\ell$; and the loss function is quadratic.

3.2 Filtering Algorithms

Under each of the sets of conditions (1) and (2) above, and the assumption that Q and σ^2 are known (see Sections 1.1, 2.2) a recursive algorithm can be obtained which provides estimates $\hat{\underset{\sim}{\theta}}_{\ell|\ell}$, successively, for $\ell=1,2,\ldots,N$. This algorithm corresponds directly to the well known Kalman filter of state estimation theory:

$$\hat{\underset{\sim}{\theta}}_k = \Phi\hat{\underset{\sim}{\theta}}_{k-1} + \frac{P_{k|k}\underset{\sim}{u}_k}{\sigma^2}(y_k - \underset{\sim}{u}_k{}^T\Phi\hat{\underset{\sim}{\theta}}_{k-1}) \qquad (3.2.1)$$

$$P_{k|k} = P_{k|k-1} - P_{k|k-1}\underset{\sim}{u}_k(\sigma^2 + \underset{\sim}{u}_k{}^T P_{k|k-1}\underset{\sim}{u}_k)^{-1}\underset{\sim}{u}_k{}^T P_{k|k-1} \qquad (3.2.2)$$

$$P_{k|k-1} = \Phi P_{k-1|k-1}\Phi^T + \Gamma Q \Gamma^T \qquad (3.2.3)$$

Here $P_{k|\ell} = E(\underset{\sim}{\theta}_k - \hat{\underset{\sim}{\theta}}_{k|\ell})(\underset{\sim}{\theta}_k - \hat{\underset{\sim}{\theta}}_{k|\ell})^T$ for $\ell = k-1,k$,

where $\hat{\underset{\sim}{\theta}}_{k|k-1} = \Phi\hat{\underset{\sim}{\theta}}_{k-1}$

We usually assume $\underset{\sim}{\theta}_0$ is a mean zero random variable (Gaussian under conditions (1)), with a large diagonal variance-covariance matrix, to indicate very little confidence in the initial estimate $\underset{\sim}{\theta}_0$ (see Section 5.3). This represents an approximately uniform prior distribution.

Many derivations of this algorithm have appeared since Kalman's original solution (Kalman, 1960) which was under condition (2) (Rauch, Tung and Striebel, 1965; Kailath and Frost, 1968; Young, 1965; 1969; Duncan and Horn, 1972). The derivation of Bryson and Ho (1969) possibly gives the most lucid solution under condition (1). These authors derive equations of evolution for the conditional mean and variance in the densities $p(\underset{\sim}{\theta}_\ell|\underset{\sim}{Y}_\ell)$. Although there are some alternative forms of this algorithm, they are very similar with respect to the criteria of computational efficiency and numerical stability.

3.3 Smoothing Algorithms

In the case of obtaining an estimate of $\underset{\sim}{\theta}_\ell$ on the basis of $\underset{\sim}{Y}_\ell$, for $k>\ell$ (the smoothing problem) under either conditions (1) or (2), the solution is not so clearly defined. Norton (1975) has examined a number of different solutions to the problem, each of which has various advantages and disadvantages in terms of the two criteria of computational efficiency and numerical stability. Nevertheless, it should be noted that theoretically they all provide the same result, and are obtainable from each other, although by fairly lengthy manipulation. Once again Q and σ^2 are assumed known.

The simplest form of the algorithm is obtained by maximizing the Gaussian density $p(\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N|Y_N)$ with respect to $\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N$ to give the conditional mean (or equivalently, the maximum *a posteriori*) estimate under condition (1). By Bayes' theorem

$$p(\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N|\underset{\sim}{Y}_N) = \frac{p(\underset{\sim}{Y}_N|\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N)p(\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N)}{p(\underset{\sim}{Y}_N)}$$

$$= \frac{\prod_{k=1}^{N} p(Y_k|\underset{\sim}{\theta}_k)\prod_{k=1}^{N} p(\underset{\sim}{\theta}_k|\underset{\sim}{\theta}_{k-1})p(\underset{\sim}{\theta}_0)}{p(\underset{\sim}{Y}_N)} \qquad (3.3.1)$$

using the Markov property of $\{\underset{\sim}{\theta}_k\}_{k=0}^{N}$.

Therefore, the maximization of this density is equivalent to maximizing

$$F = \prod_{k-1}^{N} p(Y_k|\underset{\sim}{\theta}_k)\prod_{k=1}^{N} p(\underset{\sim}{\theta}_k|\underset{\sim}{\theta}_{k-1})p(\underset{\sim}{\theta}_0), \text{ with respect to } \underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N\,. \qquad (3.3.2)$$

Now $\underset{\sim}{\theta}_k|\underset{\sim}{\theta}_{k-1}$ has density which is $N(\Phi\underset{\sim}{\theta}_{k-1},\Gamma Q\Gamma^T)$, and $Y_k|\underset{\sim}{\theta}_k$ has density which is $N(\underset{\sim}{u}_k^T\underset{\sim}{\theta}_k,\sigma^2)$. Therefore the sum of the quadratic forms in the exponents of the densities in (3.3.2) is

$$J = \frac{1}{\sigma^2}\sum_{k=1}^{N}(y_k-\underset{\sim}{u}_k^T\underset{\sim}{\theta}_k)^2 + \sum_{k=1}^{N}(\underset{\sim}{\theta}_k-\Phi\underset{\sim}{\theta}_{k-1})^T(\Gamma Q\Gamma^T)^{-1}(\underset{\sim}{\theta}_k-\Phi\underset{\sim}{\theta}_{k-1})$$

$$+ (\underset{\sim}{\theta}_0-\hat{\underset{\sim}{\theta}}_0)^T P_0^{-1}(\underset{\sim}{\theta}_0-\hat{\underset{\sim}{\theta}}_0) \qquad (3.3.3)$$

where $\hat{\underset{\sim}{\theta}}_0$ and P_0 are the initial estimates as in the filtering case.

Typically

$$\hat{\underset{\sim}{\theta}}_0 = 0, \; P_0 = 10^6 I_m.$$

J can be minimized by differentiating with respect to $\underset{\sim}{\theta}_k$ and $\underset{\sim}{w}_k$, where the constraint $\underset{\sim}{\theta}_k = \Phi\underset{\sim}{\theta}_{k-1} + \Gamma\underset{\sim}{w}_k \quad k = 1,2,\ldots,N$ is introduced via Lagrange multipliers $\underset{\sim}{\lambda}_k$, $k=1,2,\ldots,N$. Then we have to differentiate

$$\begin{aligned} J^0 = {} & \frac{1}{2\sigma^2} \sum_{k=1}^{N} (\underset{\sim}{y}_k - \underset{\sim}{u}_k{}^T \underset{\sim}{\theta}_k) + \frac{1}{2} \sum_{k=1}^{N-1} \underset{\sim}{w}_k{}^T Q^{-1} \underset{\sim}{w}_k \\ & + \frac{1}{2} (\underset{\sim}{\theta}_0 - \hat{\underset{\sim}{\theta}}_0)^T P_0{}^{-1} (\underset{\sim}{\theta}_0 - \hat{\underset{\sim}{\theta}}_0) \\ & + \sum_{k=1}^{N-1} \underset{\sim}{\lambda}_k{}^T (\underset{\sim}{\theta}_{k+1} - \Phi\underset{\sim}{\theta}_k - \Gamma w_k) \end{aligned} \qquad (3.3.4)$$

with respect to $\underset{\sim}{w}_k, \underset{\sim}{\theta}_k, \underset{\sim}{\lambda}_k$.

Doing this results in the equations

$$\underset{\sim}{\theta}_{k+1|N} = \Phi\hat{\underset{\sim}{\theta}}_{k|N} - \Gamma Q \Gamma^T \underset{\sim}{\lambda}_k \qquad (3.3.5)$$

$$\underset{\sim}{\lambda}_k = \Phi\underset{\sim}{\lambda}_{k+1} \frac{\underset{\sim}{u}_k}{\sigma^2} (y_{k+1} - \underset{\sim}{u}_k{}^T \hat{\underset{\sim}{\theta}}_{k+1|N}) \quad k = 0,1,\ldots,N-1 \qquad (3.3.6)$$

$$\underset{\sim}{\lambda}_{-1} = P_0{}^{-1} (\hat{\underset{\sim}{\theta}}_0 - \hat{\underset{\sim}{\theta}}_{0|N}) \qquad (3.3.7)$$

$$\underset{\sim}{\lambda}_N = \underset{\sim}{0} \qquad (3.3.8)$$

These equations constitute a two-point boundary value problem, with split initial conditions (3.3.7) and (3.3.8). We can solve the problem by obtaining $\hat{\underset{\sim}{\theta}}_{N|N}$ from a filtering run as described in Section 3.2 to give terminal conditions on both $\underset{\sim}{\lambda}_N$ and $\hat{\underset{\sim}{\theta}}_{N|N}$, and hence solve the equation backwards in time. Norton shows, however, that the resulting algorithm is potentially numerically unstable, by writing the solution in the form

$$\begin{pmatrix} \underset{\sim}{\lambda}_k \\ \\ \hat{\underset{\sim}{\theta}}_{k|N} \end{pmatrix} = P \begin{pmatrix} \underset{\sim}{\lambda}_{k+1} \\ \\ \hat{\underset{\sim}{\theta}}_{k+1|N} \end{pmatrix} + \underset{\sim}{t}_k,$$

where t_k does not involve $\underset{\sim}{\lambda}_k$ or $\underset{\sim}{\theta}_{k|N}$. Then it can be shown that P has eigenvalues outside the unit circle.

Rauch, Tung and Striebel (1965) maximize the marginal density $p(\underset{\sim}{\theta}_k, \underset{\sim}{\theta}_{k+1} | \underset{\sim}{Y}_N)$ with respect to $\underset{\sim}{\theta}_k, \underset{\sim}{\theta}_{k+1}$. This is equivalent to the first procedure, since the random variables $\underset{\sim}{\theta}_k, \underset{\sim}{\theta}_{k+1}$, conditional on $\underset{\sim}{Y}_N$ have expected value equal to the corresponding part of the expectation of $\underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_N$ conditional on $\underset{\sim}{Y}_N$, and these expectations maximize the corresponding densities. Manipulating the densities once again yields a quadratic form to be minimized, and the resulting algorithm is

$$\hat{\underset{\sim}{\theta}}_{k|N} = \hat{\underset{\sim}{\theta}}_{k|k} + P_{k|k}\Phi^T P^{-1}_{k+1|k} + (\hat{\underset{\sim}{\theta}}_{k+1|N} - \Phi\hat{\underset{\sim}{\theta}}_{k|k}) \quad (3.3.9)$$

with notation as in equations (3.2.1), (3.2.2), (3.2.3). This form avoids the use of the adjoint variable $\underset{\sim}{\lambda}_k$, but introduces the numerical complications of inverting $P_{k+1|k}$ at each step. The storage requirements of this algorithm are also higher, because $\hat{\underset{\sim}{\theta}}_{k|k}$, $P_{k|k}$ and $P_{k+1|k}$ need to be saved from the filtering run.

Norton concludes that the most useful form of the smoothing algorithm in this case is that derived by Bryson and Ho (1975) under condition (1). The algorithm can also be derived from the general form of the smoothing solution under condition (2), given in equation (2.3.4). Evaluating the covariances in this equation, and defining the variable $\underset{\sim}{\lambda}_k$ recursively by

$$\underset{\sim}{\lambda}_k = (I_m - P_{k+1|k+1} \frac{\underset{\sim}{u}_k \underset{\sim}{u}_k^T}{\sigma^2})^T \, (\Phi\underset{\sim}{\lambda}_{k+1} - \frac{\underset{\sim}{u}_k}{\sigma^2} (y_k - \underset{\sim}{u}_k^T \Phi \underset{\sim}{\hat{\theta}}_{k|k}) \qquad (3.3.10)$$

$$\underset{\sim}{\lambda}_N = \underset{\sim}{0}$$

we can obtain the smoothed estimates recursively backwards either from

$$\underset{\sim}{\hat{\theta}}_{k|N} = \underset{\sim}{\hat{\theta}}_{k|k} - P_{k|k}\Phi_k^T \underset{\sim}{\lambda}_k \qquad (3.3.11)$$

or

$$\underset{\sim}{\hat{\theta}}_{k|N} = \Phi^{-1}(\underset{\sim}{\hat{\theta}}_{k+1|N} + \Gamma Q \Gamma^T \underset{\sim}{\lambda}_k) \qquad (3.3.12)$$

Norton shows that in this case, the backward recursion is stable.

Other derivations are also considered by Norton, but are rejected because they provide algorithms which either involve matrix inversion or require greater storage space than the algorithms given above by (3.3.10), and either (3.3.11) or (3.3.12).

The variance-covariance matrix of the error $\underset{\sim}{\theta}_k - \underset{\sim}{\hat{\theta}}_{k|N}$ in the smoothed estimate can also be obtained in a number of ways. Rauch, Tung and Striebel (1965) give

$$P_{k|N} = P_{k|k} + P_{k|k}\Phi P_{k+1|k}^{-1}(P_{k+1|N} - P_{k+1|k})P_{k+1|k}^{-1}\Phi P_{k|k};$$

Bryson and Ho (1975) avoid the matrix inversion with a slightly lengthier algorithm. In general, however, it is not essential to compute this covariance, since, unlike in the case of the filtering algorithm, it is not needed to generate the parameter estimate $\hat{\theta}_{k|N}$. Of course, this will mean that the exact error covariance properties of the smoothed estimate will not be available to the analyst. However, since $P_{k|N}$ is bounded above by $P_{k|k}$, it may well be that if $P_{k|k}$ is "small enough", then this will be sufficient information for most practical purposes.

It should be noted that in all the algorithms in this chapter, all the matrices, $P_{k|\ell}$, $\ell = k,k-1$, $k=1,2,\ldots,N$, and Q can be divided through by σ^2 as a normalizing factor, and the algorithms when processed using the normalized form. This eliminates the need for σ^2, but of course $P_{k|\ell}$ will not then be the true error variance-covariance matrices.

CHAPTER 4 : ALGORITHMS FOR ESTIMATION IN MODEL II

4.1 Introduction

Once again we refer to equations (2.3.1), (2.3.2). For model II, we have

$$y_k = \frac{B_k}{A_k} u_k + e_k \qquad (4.1.1)$$

$$\underset{\sim}{\theta}_k = \Phi\underset{\sim}{\theta}_{k-1} + \Gamma\underset{\sim}{\nu}_k \qquad (4.1.2)$$

In this case, the relationship between $\underset{\sim}{\theta}_k$ and e_k, $k=1,2,\ldots,N$, is not linear. Therefore under condition (1) of Chapter 3, while $p(e_k)$ is still Gaussian, not all conditional densities are now necessarily Gaussian. This can be clearly illustrated by taking a simple case.

If $B_k(z^{-1}) \overset{\Delta}{=} b_{OK} = b_k$; and $A_k(z^{-1}) = 1 + a_{1k}z^{-1} \overset{\Delta}{=} 1 + a_k z^{-1}$,

then $y_{k+1} = b_{k+1}u_{k-1} - a_{k+1}x_k + \varepsilon_{k+1}$

$$= b_{k+1}u_{k+1} - a_{k+1}(b_k u_k - a_k x_{k-1}) + \varepsilon_{k+1}$$

Therefore the density $p(\underset{\sim}{Y}_k)$, being the sum of random variables some of which are products of Gaussian random variables, is not itself Gaussian. Hence conditional means and variances cannot be obtained so easily. Procedures based on condition (2) also encounter difficulty because of the non-linearity : the quantities $E(\underset{\sim}{\theta}_k\varepsilon_t)$, $E(\varepsilon_t^2)$ in (2.3.4) cannot be evaluated easily as they could be for model I.

It can be seen that there is not one general procedure for deriving algorithms to estimate $\underset{\sim}{\theta}_k$ in this case. Moreover, not all methods produce the same estimate, as was the situation in the previous chapter. A large number of estimation algorithms have been employed in general non-linear state estimation problems. For example Sorenson and Stubberud (1968) obtain (approximate) equations of evolution for conditional means and variances by assuming that the conditional densities at each time point are Gaussian, and computing their means and variances. The context under consideration is one where second-order non-linearities are the only non-negligible higher order effects. Another general solution can be obtained under condition (1), with the required estimate being the maximum *a posteriori* estimate. Then following Cox (1964) we can proceed from an equation analogous to (3.3.1), to obtain

$$p(\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_N|\underset{\sim}{Y}_N) = \frac{\prod_{k=1}^{N} p(Y_k|\underset{\sim}{\theta}_k)\prod_{k=1}^{N} p(\underset{\sim}{\theta}_k|\underset{\sim}{\theta}_{k-1})p(\underset{\sim}{\theta}_0)}{p(\underset{\sim}{Y}_N)}$$

Here, once again assuming Q, σ^2 known, the exponent in the densities of interest is

$$J = \frac{1}{2\sigma^2}\sum_{k=1}^{N}(y_k-x_k(\underset{\sim}{\theta}_k))^2 + \frac{1}{2}\sum_{k=1}^{N}(\underset{\sim}{\theta}_k - \Phi\underset{\sim}{\theta}_{k-1})^T(\Gamma Q\Gamma^T)^{-1}(\underset{\sim}{\theta}_k - \Phi\underset{\sim}{\theta}_{k-1})$$

$$+ \frac{1}{2}(\underset{\sim}{\theta}_0-\hat{\underset{\sim}{\theta}}_0)P_0^{-1}(\underset{\sim}{\theta}_0-\hat{\underset{\sim}{\theta}}_0) \qquad (4.1.3)$$

where, for model II

$$x_k(\underset{\sim}{\theta}_k) \overset{\Delta}{=} \frac{B_k(z^{-1})}{A_k(z^{-1})} u_k, \qquad (4.1.4)$$

and P_0, $\underset{\sim}{\theta}_0$ are defined as in (3.3.3).

The minimization of J with respect to $\underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_N$ can be accomplished by introducing the Lagrange multipliers as before in (3.3.4) to convert the problem into one of minimizing

$$J^o = \frac{1}{2\sigma^2} \sum_{k=1}^{N} (y_k - x_k(\underset{\sim}{\theta}_k))^2 + \frac{1}{2} \sum_{k=1}^{N} \underset{\sim}{w}_k^T Q^{-1} \underset{\sim}{w}_k$$

$$+ \frac{1}{2} (\underset{\sim}{\theta}_0 - \hat{\underset{\sim}{\theta}}_0) P_0^{-1} (\underset{\sim}{\theta}_0 - \hat{\underset{\sim}{\theta}}_0)$$

$$+ \sum_{k=1}^{N-1} \underset{\sim}{\lambda}_k^T (\underset{\sim}{\theta}_{k+1} - \Phi \underset{\sim}{\theta}_k - \Gamma \underset{\sim}{w}_k)$$

with respect to $\underset{\sim}{\lambda}_k, \underset{\sim}{\theta}_k$, k=0,1,...,N and $\underset{\sim}{w}_k$, k=1,2,...,N. Setting the derivative of J^o with respect to these quantities equal to zero gives the discrete non-linear two point boundary value problem[†]

$$\hat{\underset{\sim}{\theta}}_{k+1|N} = \Phi \hat{\underset{\sim}{\theta}}_{k|N} - \Gamma Q \Gamma^T \underset{\sim}{\lambda}_k \qquad (4.1.5)$$

$$\underset{\sim}{\lambda}_k = \Phi \underset{\sim}{\lambda}_{k+1} - \left(\frac{\partial x_k}{\partial \underset{\sim}{\theta}_k}\bigg|_{\underset{\sim}{\theta}_k = \hat{\underset{\sim}{\theta}}_{k+1|N}}\right)^T \frac{1}{\sigma^2} (y_{k+1} - x_k(\hat{\underset{\sim}{\theta}}_{k+1|N})) \qquad (4.1.6)$$

with boundary conditions on $\hat{\underset{\sim}{\theta}}_{0|N}$ and $\underset{\sim}{\lambda}_N$.

[†] A similar two point boundary value problem can be obtained by applying the discrete maximum principle (Sage and Melsa, 1971) to (4.1.3).

It is not possible to convert (4.1.5)-(4.1.6) into a one-sided boundary value problem by obtaining $\underset{\sim}{\theta}_{N|N}$ from a filtering run, as was done to solve (3.3.5)-(3.3.8). This is because the filtering maximum *a posteriori* solution cannot be obtained in closed form. There is a vast armoury of numerical techniques available to solve the boundary value problem, but they are cumbersome and do not guarantee a solution, particularly when a good initial estimate is not available (Sage and Melsa, 1971^2). Sage and Ewing (1970) demonstrate one example of such a procedure.

Because of these difficulties in obtaining algorithms for model II, we now turn to examine procedures which take advantage of the special nature of the non-linearity in (4.1.1). We continue to assume Q and σ^2 known.

4.2 Least Squares Estimation

Although it is true that many of the estimation methods described in this thesis can be placed in a least squares context, the term is used here to refer to the approximating of a correlated sequence of random variables by an i.i.d. sequence. If we write (4.1.1) as

$$A_k y_k = B_k u_k + A_k e_k \qquad (4.2.1)$$

then $$y_k = B_k u_k - (A_k - 1) y_k + e_k^* , \qquad (4.2.2)$$

where $$e_k^* = A_k e_k .$$

Equation (4.1.2) now provides a form for the problem, such that if e_k^* is assumed i.i.d., with mean zero and variance σ^2, the filtering and smoothing algorithms from Chapter 3 can be applied, with $\underset{\sim}{u}_k^T = (-y_{k-1},\ldots,-y_{k-n},u_k,\ldots,u_{k-n})$. For negative j, u_j and y_j may be taken as zero. This will be discussed in more detail in Section 5.2. The disadvantages of this scheme is that biased estimates of the parameter values may result (see Section 6.1).

4.3 Extended Least Squares Estimation

This procedure is used by Norton (1975) in the estimation of $\underset{\sim}{\theta}_k$ in the model

$$A_k y_k = B_k u_k + C_k e_k \qquad (4.3.1)$$

Here all quantities are defined as in Section 1.1, with

$$C_k = C_k(z^{-1}) \overset{\Delta}{=} 1 + c_{1k}z^{-1} +,\ldots,+ c_{nk}z^{-n}$$

The parameter evolution equation is as in Section 2.2, with $\underset{\sim}{\theta}_k$ now defined as

$$\underset{\sim}{\theta}_k^T = (a_{1k},\ldots,a_{nk},b_{0k},\ldots,b_{nk},c_{1k},\ldots,c_{nk})$$

Applying Norton's method to (4.2.1) there is some redundancy, since the parameters $a_{1k},\ldots,a_{nk}$ are estimated twice. The concept, however, can still be used. Rewriting (4.3.1) as

$$y_k = B_k u_k + (A_k-1)y_k + (C_k-1)e_k + e_k$$

the model is once again in a form where the algorithms of Chapter 3 can be applied, except that the terms $(C_k-1)e_k$ involves the unknown noise terms $e_{k-1}, e_{k-2}, \ldots, e_{k-n}$. These terms can, nevertheless, be estimated recursively via

$$\hat{e}_{k-r} = y_{k-r} - \hat{\underset{\sim}{u}}_{k-r}\hat{\underset{\sim}{\theta}}_{k-r|k-r}$$

where now

$$\underset{\sim}{u}_{k-r}^T = (-y_{k-r-1}, \ldots, -y_{k-r-n}, u_{k-r}, \ldots, u_{k-r-n}, \hat{e}_{k-r-1}, \ldots, \hat{e}_{k-r-n}).$$

Noise terms with negative indices are taken at their mean value, zero, or can be estimated in other ways (see Section 5.2).

This procedure corresponds, in the case of constant parameters, to the RELS algorithm of Soderström et al., (1974) or the AML algorithm of Young et al.,(1971), or Young (1974), who also use the method with time-varying parameters. While it is quite satisfactory in many situations, difficulties may arise. These may be due firstly to the abovementioned redundancy arising in the model under consideration here, and secondly to possible large inaccuracies in early noise estimates.

4.4 Instrumental Variable Estimation

We once again rewrite the equation (4.1.1), this time in the form

$$y_k = B_k u_k - (A_k-1)x_k + e_k \tag{4.4.1}$$

where

$$x_k = \frac{B_k}{A_k} u_k, \text{ as before.}$$

Then, if x_k, the noise free output, were known, the model would once again be in a form where the algorithms of Chapter 3 could be applied. Now since estimates of A_k and B_k, $k = 1,2,\ldots,N$, can be provided by either least squares (Section 4.2) or extended least squares (Section 4.3), it is possible to also estimate x_k, via

$$\hat{x}_k = \frac{\hat{B}_k}{\hat{A}_k} u_k$$

Substituting in (4.4.1), this gives an observation equation of the form

$$y_k = B_k u_k - (A_k - 1)\hat{x}_k + e_k.$$

We can estimate A_k and B_k again from this equation, taking

$$\underset{\sim}{u}_k^T = (-\hat{x}_{k-1},\ldots,-\hat{x}_{k-n},u_k,\ldots,u_{k-n}) \qquad (4.4.2)$$

This procedure can be iterated until there is no significant change in the estimates.

It would be difficult to justify this analytically, and indeed, there is no guarantee that the iterative procedure would even improve estimates. However, the method is closely related to that of Young (1969; 1974), which is developed in the instrumental variables framework. Considering first the constant parameter situation, we can write, as before,

$$y_k = Bu_k - (A-1)y_k + e_k^* \qquad (4.4.3)$$

where $e_k^* = A\ e_k$. It is known that the least squares estimate of A and B in (4.4.3) is biased, due to the correlation between e_k^* and y_k, but that the use of an instrumental variable can remove this problem (Kendall and Stuart, 1961). The instrumental variable vector chosen by Young et al., (1971) is

$$\underset{\sim}{x}_k{}^T = (-\hat{x}_{k-1},\ldots,-\hat{x}_{k-n},u_k,\ldots,u_{k-n})$$

which satisfies the criteria of being highly correlated with $\underset{\sim}{x}_k{}^T = (x_{k-1},\ldots,x_{k-n},u_k,\ldots,u_{k-n})$ while being uncorrelated with e_k^*. The estimates $\hat{x}_j$ are obtained from a previous estimate of the parameters, via $\hat{x}_j = (\hat{B}/\hat{A})u_j$. The resulting (recursive instrumental variable) algorithm is

$$\hat{\underset{\sim}{\theta}}_k = \hat{\theta}_{k-1} + P_{k-1}\hat{\underset{\sim}{x}}_k\{1 + \underset{\sim}{z}_k{}^T P_{k-1}\hat{x}_k\}^{-1}\{y_k - \underset{\sim}{z}_k{}^T\hat{\underset{\sim}{\theta}}_{k-1}\} \qquad (4.4.4)$$

$$P_k = P_{k-1} - P_{k-1}\hat{\underset{\sim}{x}}_k\{1 + \underset{\sim}{z}_k{}^T P_{k-1}\hat{\underset{\sim}{x}}_k\}^{-1}\underset{\sim}{z}_k{}^T P_{k-1} \qquad (4.4.5)$$

where

$$\underset{\sim}{z}_k{}^T = (-y_{k-1},\ldots,-y_{k-n},u_k,\ldots,u_{k-n}),$$

and the procedure is iterative as before.

The extension of the algorithm (4.4.4)-(4.4.5) to time varying parameters is made in Young (1965; 1969) with the resulting algorithms differing from the filtering form of the algorithm outlined above only in that $\underset{\sim}{z}_k$ is replaced by $\hat{\underset{\sim}{x}}_k$ throughout. The method of updating the auxiliary model $\hat{B}_k/\hat{A}_k$ also differs. In Young (1969), the auxiliary model is kept constant during each iteration†. In the formulation

† This was due to limitations on the analog equipment used in the hybrid (analog-digital) mechanisation of the corresponding algorithm in the estimation of differential equation models.

implemented as above, the auxiliary model is taken from the smoothed estimate obtained in the previous iteration, for each time point. It should nevertheless be pointed out that quite reasonable results can be obtained, in many cases, if the auxiliary model remains constant (see Young, 1969).

4.5 Refined Instrumental Variable Estimation

One of the most frequently applied methods of overcoming problems of non-linearity in state estimation contexts is through the use of a linearization of the observation and system equations about some reference trajectory, which may either be a successively updated state estimate or an appropriate estimate obtained by some other means. For example, we could obtain an approximate (filtered) estimate in model II by proceeding as follows. Under condition (2) of Chapter 3, we can write, in a similar manner to (2.3.4),

$$\hat{\underset{\sim}{\theta}}_{k|k} = \sum_{t=1}^{k} \{E(\hat{\underset{\sim}{\theta}}_k \varepsilon_t)/E(\varepsilon_t{}^2)\} \varepsilon_t$$

where ε_t are the linear innovations defined in (2.3.3). This can be expressed recursively as :

$$\hat{\underset{\sim}{\theta}}_{k|k} = \hat{\underset{\sim}{\theta}}_{k|k-1} + \{E(\underset{\sim}{\theta}_k \varepsilon_k)/E(\varepsilon_k{}^2) \varepsilon_k \qquad (4.5.1)$$

The covariance quantities in (4.5.1) cannot be easily evaluated exactly for model II, so that the innovations cannot be obtained exactly either. It is possible to approximate ε_k by

$\varepsilon_k \simeq y_k - x_k(\Phi\hat{\underset{\sim}{\theta}}_{k-1|k-1})$, where $x_k(.)$ is as in (4.1.4). Also put $\hat{\underset{\sim}{\theta}}_{k|k-1} = \Phi\hat{\underset{\sim}{\theta}}_{k-1|k-1}$. To obtain the covariances we take a first order Taylor expansion of $x_k(\underset{\sim}{\theta}_k)$ about $\hat{\underset{\sim}{\theta}}_{k|k-1}$:

$$x_k(\underset{\sim}{\theta}_k) \simeq x_k(\hat{\underset{\sim}{\theta}}_{k|k-1}) + H_k^T(\underset{\sim}{\theta}_k - \hat{\underset{\sim}{\theta}}_{k|k-1})$$

where $H_k = \frac{\partial x_k}{\partial \underset{\sim}{\theta}_k}\Big|_{\underset{\sim}{\theta}_k = \hat{\underset{\sim}{\theta}}_{k|k-1}}$. We can then obtain approximations to the expressions in (4.5.1) which yield a filtering algorithm identical to (3.2.1)-(3.2.3) except that $\underset{\sim}{u}_k$ is replaced by H_k. It should be noted that $P_{k|\ell}$, $\ell = k-1,k$ are no longer true variance-covariance matrices, but approximations. The approximate smoothing solution can also be obtained from the algorithms of Section 3.3, with $\underset{\sim}{u}_k$ replaced by H_k.

Upon examination, it can be seen that

$$H_k^T = \left(-\frac{\hat{B}_{k|k-1}}{\hat{A}^2_{k|k-1}}\, u_{k-1}, \ldots, -\frac{\hat{B}_{k|k-1}}{\hat{A}^2_{k|k-1}}\, u_{k-n}, \frac{1}{\hat{A}_{k|k-1}}\, u_k, \ldots, \frac{1}{\hat{A}_{k|k-1}}\, u_{k-n}\right)$$

In the terminology of Young (1976) H_k is therefore a vector of pre-filtered variables, as compared with the unfiltered variables given by (4.4.2) in the instrumental variable algorithm. The algorithm corresponds, in the constant parameter case, to a smoothing version of the symmetric form of the refined IV algorithm (Young and Jakeman, 1978). It can also be compared with the RML algorithm of Soderström et al., (1974), in which a similar linearization produces an algorithm for estimating the parameters in (4.3.1). A form corresponding to the asymmetric refined IV algorithm (Young, 1976) can also be derived.

Once again, it would be difficult to theoretically substantiate any claims of increased benefit gained from this refined algorithm, as compared with the instrumental variable form of Section 4.3. However it has been demonstrated by Young and Jakeman (1978) in simulations, and by Solo (1978) in a plausibility argument, that the refined form produces asymptotically efficient estimates of constant A and B parameters, and there is often a clear reduction in estimation error variance to be gained over the IV algorithm. Therefore, when the parameters are varying in a manner closely approximating a random walk, improved performance may be gained from the refined form.

It should be noted that in practice, the pre-filters and auxiliary model would not be updated at each step. Rather they would be given by a previous estimation run as is done with the auxiliary model in the instrumental variable form. This eliminates stability problems which have been found to occur in the fully recursive form for constant parameters (Young and Jakeman, 1978), and so would presumably be even more likely to occur in the varying parameter situation. An iterative procedure as in the recursive IV case could also be applied here.

4.6 Conclusion

While we have not obtained a definite solution to the problem of estimating time varying parameters for model II (assuming Q and σ^2 known), it has been shown that for this purpose there are a number of satisfactory approximations which can be applied. Most of these relate to methods used extensively for estimating constant parameters, and avoid the complications which may be encountered when applying general non-linear state estimation algorithms.

CHAPTER 5 : UTILIZATION OF THE ALGORITHMS

5.1 Introduction

Thus far we have considered means by which one might model parametric variation in the models I and II, and estimate parameters in these models. There remain, however, some difficulties to be overcome in the practical implementation of the algorithms we have obtained. In Section 5.2, the process which may lead to the adoption of a time-varying parameter model is discussed. This can be thought of as 'identification of structure', in the sense of Box and Jenkins (1970). Section 5.3 is concerned with ways of obtaining values of the *program parameters* . These are the variances, Q and σ^2, and the initial conditions $\underset{\sim}{\theta}_0, P_{0|0}$, (as defined in Chapters 2, 3 and 4) which have been so far assumed known. Finally in Section 5.4, some asymptotic properties of the estimation procedures are considered.

5.2 Identification of Time-Varying Structure

We can recognize three possible stages in the process of adopting a time-varying parameter model. Firstly, examination of constant parameter results; secondly, hypothesis testing concerning the possibility of parametric change, and thirdly, the estimation of a time varying parameter model. The third stage has been considered in some detail already, so we will here briefly consider some aspects of the first two stages.

5.2.1 Examination of constant parameter results

The use of recursive estimation methods in constant parameter time series and regression models has come into favour recently (Young, 1974; Soderström et al., 1974). Not only have they been found to provide computationally attractive means of obtaining consistent, efficient, parameter estimates (Young, 1976), but also covergence characteristics can be conveniently examined by reference to graphical outputs of the recursive parameter estimates. In this way, it is possible to ascertain whether the estimates are slow in converging, or if, indeed they fail to converge.

Slow convergence or failure to converge can occur for a number of reasons. Firstly, there could be an identifiability† problem associated with the model. In the case of model I, this could arise through multicollinearity of the inputs (regressors in this case) $u_k^{(i)}$, = 1,2,...,M. Tests to detect this, such as the multiple correlation test, are well known (Kendall and Stuart, 1961). Multicollinearity is manifested in near-singularity of the information matrix U^TU, where

$$U = \begin{pmatrix} u_1^{(1)} & \cdots & u_1^{(m)} \\ u_N^{(1)} & \cdots & u_N^{(m)} \end{pmatrix}$$

†See Hannan (1971) for a general discussion of identifiability.

which leads to a high (normalized) estimation error covariance matrix S_N. In model II, an identifiability problem could arise through pole-zero cancellation in the transfer function $\frac{B}{A}$, indicating that a model of too high an order is being fitted to the data. Once again, this is manifested in a large estimation error covariance matrix, and a number of procedures can be used to test whether this is the case (Young et al., 1978). Again identifiability problems can arise because the input signal u_k is not 'sufficiently exciting' (Aström and Bohlin, 1966). For example, a second order system is not identifiable when perturbed only by a single sinusoidal input : at least two different frequency components are required to avoid identifiability problems (see Young et al., 1971).

If the possibility of non-identifiability has been eliminated, then the reason for slow convergence of the parameters is that a single model is not appropriate at all time points, and that there appears to be some variation in the parameters. An examination of plotted residuals (Draper and Smith, 1967; for model I) or innovations (Harvey and Phillips, 1976, for model II) in a constant parameter model, may also corroborate evidence of this kind, since certain types of parametric variation may appear as a systematic component in residuals or innovations. If there is such evidence of parametric variation, then we may proceed to the second stage outlined above, provided the indicated variance appears to be physically meaningful.

5.2.2 Testing the hypothesis of parametric change

This stage in the procedure outlined at the start of this section is not considered by the author to be essential in the context of the present work. In situations where the methods of this thesis may be applied, we are concerned with examining the plausibility of some types of parameter variation, by reference to the results obtained in the estimation, *in conjunction with physical knowledge of the system being studied.* Therefore, while it may be claimed that an assertion concerning a statistical model must be accompanied by an appropriate test of statistical significance, it is considered that the 'positive or negative' result obtained from a hypothesis test may be too restrictive to be generally useful. Nevertheless, various authors have discussed methods of carrying out a formal hypothesis test concerning parametric charge, and we make brief mention of some of these here.

Brown et al. (1975) appear to have suggested the first test for general non-constancy of parameters in model I, the regression model; they derive approximate distributions for the sum of, and sum of squares of, recursive residuals (or filtered innovations, in our terminology), under the null hypothesis of constant parameters. For the same model, Garbade (1977) suggests using a likelihood ratio test of the null hypothesis $Q = 0$ against the alternative $Q \neq 0$, with Q as in Section 2.3, and taking $\Phi = \Gamma = I_m$ in (2.3.2).

He then goes on to compare the two tests, using simulations of three different types of parametric change in a simple regression model. The latter test is shown to be superior in a number of respects.

More recently, Pagan (1978) and Salmon (1978) have suggested the use of a Lagrange multiplier test. However, there are no studies as yet available to demonstrate the use of this test in practice, or to compare it with other hypothesis tests in this context.

5.3 The Choice of Program Parameters

In order to implement the algorithms of Chapters 3 and 4, it is necessary to choose values of the program parameters mentioned in Section 5.1. It will become clear that there is a certain amount of freedom associated with the values of these parameters. Nevertheless, it is useful both to understand the effect of using different values of these parameters, and to have an analytic method of choosing values, should this be called for.

5.3.1 Parameter variance Q and measurement variance σ^2

Here, the corresponding state estimation problem, (that is, one of obtaining values of system and measurement noise levels in order to implement a filtering or smoothing algorithm) has received attention in recent years (Mehra, 1971; Neethling and Young, 1974, among others). However, no solution could be claimed as generally appropriate in the state estimation context. The respective advantages and disadvantages of some of the solutions are discussed by Neethling (1974), and, in a parameter estimation context, by Bennett (1976). Most of these methods are aimed at the estimation of the values of Q and σ^2 (or Q and R, a matrix, in multi-output estimation situations) concurrently with the estimation of the state variables; that is, *adaptive* estimation of Q and σ^2 (or Q and R). In the context of time-variable parameter estimation, such a procedure would be neither necessary nor appropriate. The Q matrix does not have a *physical* interpretation, as it does in the state estimation problem. In the context being considered here, it may be thought of as a quantification of the expected rate of parameter variation between samples, so that when using an adaptive procedure for estimating Q, it would clearly be hard to distinguish between changes in Q and changes in the parameters themselves. It is also noteworthy that the methods of Mehra (1971), Neethling (1974) and others rely on the assumption that the process being estimated is in steady state, so that asymptotic values of the covariance matrices $P_{k|\ell}$, $\ell = k-1,k$, have been attained. As we shall see in Section 5.4, it is not possible to obtain these asymptotic values in the case of models I and II without placing further assumptions on the processes involved.

There is also a difficulty in using constant Q and σ^2 for an estimation run. This difficulty arises when the true parameter variation is not actually a random walk, and the rate of variation changes markedly during the observation period. Then the value of Q which is appropriate for one portion of the data may tend to exaggerate parameter variation in another part of the data where the variation is smaller, due to observation noise effects. Conversely, if a Q matrix is used which is appropriate for the portion of the data where the variation is smaller, the section of larger variation will be obscured, because the estimation procedures will consider that this is merely a noise effect, and therefore 'smooth' the estimate too much. This problem is particularly marked when the situation is one of trying to detect a step change in a parameter, particularly if the step is quite small in relation to the sample size N (see Section 6.1). Then a Q matrix which is able to accommodate the step adequately will amplify observation noise effects on the section where the parameter is constant, while the use of a Q matrix which estimates smoothly over the constant section will track the step slowly, and will not indicate its size or position very clearly.

As was mentioned in Section 3.3, σ^2 can be removed from all the algorithms obtained in Chapters 3 and 4, resulting in normalized variance-covariance matrices (or approximations to these in model II). This reduces the problem of choosing appropriate Q and σ^2 to one of obtaining the value of

$\hat{W} = Q/\sigma^2$, which is most appropriate in some sense. The effect of different values of Q and σ^2 can be illustrated by using the following simple model : take model I with

$$\begin{aligned} & m = 1,\ u_k = 1,\ k=1,2,\ldots,N \\ & \Phi = \Gamma = 1 \end{aligned} \tag{5.3.1}$$

(referring to equation (2.3.1),(2.3.2)).

Then the filtered estimate of the parameter θ_k is obtained from (3.1.1)-(3.1.3.) as

$$\hat{\theta}_{k|k} = \hat{\theta}_{k-1|k-1} + \frac{P_{k|k}}{\sigma^2}(y_k - \hat{\theta}_{k-1|k-1})$$

$$P_{k|k} = \frac{\sigma^2 P_{k|k-1}}{\sigma^2 + P_{k|k-1}}$$

$$= \frac{\sigma^2 P_{k-1|k-1} + Q}{\sigma^2 + P_{k-1|k-1} + Q}$$

Therefore, the weight (the 'Kalman gain') given to the innovation or one-step ahead prediction error $y_k - \hat{\theta}_{k|k-1}$ is

$$K_k = \frac{S_{k-1|k-1} + W}{1 + S_{k-1|k-1} + W} \tag{5.3.2}$$

where $S_{k-1|k-1} = P_{k-1|k-1} / \sigma^2$

It can be seen from (5.3.2) that K_k is a strictly monotonically increasing function of W; that is, strictly monotonically increasing in Q, and strictly monotonically decreasing in σ^2. This confirms the intuitive notions regarding the use of Q, discussed above. It is also of note that $K_k \to 1$ as $W \to \infty$. Thus above a certain level, large changes in the value of W used do not affect the estimation greatly. Also,

$$K_k \to \frac{S_{k-1|k-1}}{1 + S_{k-1|k-1}} \quad \text{as } W \to 0$$

which corresponds, in the limit (W = 0), to the constant parameter recursive least squares estimator (Plackett, 1950).

For the smoothed estimate in this model, the algorithm (3.3.9) yields

$$\begin{aligned}\hat{\theta}_{k|N} &= \hat{\theta}_{k|k} + P_{k|k}P_{k+1|k}^{-1}(\hat{\theta}_{k+1|N} - \hat{\theta}_{k|k}) \\ &= \hat{\theta}_{k|k} + \frac{P_{k|k}}{P_{k|k} + Q}(\hat{\theta}_{k+1|N} - \hat{\theta}_{k|k})\end{aligned}$$

From this, it can be seen that the smoothing procedure adjusts the filtered estimate at time k by the weighted difference between the one step ahead prediction from $\hat{\theta}_{k|k}$ and the smoothed estimate at time k+1. The weighting here is a strictly monotonically decreasing function of Q. This indicates that for large values of Q, the adjustment obtained by smoothing is small, so that the smoothed estimate 'follows' the filtered estimate closely. On the other hand,

$$\frac{P_{k|k}}{P_{k|k} + Q} \to 1 \quad \text{as } Q \to 0$$

so that, for Q = 0, the constant parameter situation,

$$\begin{aligned}\hat{\theta}_{k|N} &= \hat{\theta}_{k|k} + (\hat{\theta}_{k+1|N} - \hat{\theta}_{k|k}) \\ &= \hat{\theta}_{k+1|N} \\ &= \hat{\theta}_{N|N}\end{aligned}$$

Thus, as expected, for a parameter assumed constant, the smoothed estimate is constant over the observation period, and equal to the final filtered estimate $\hat{\theta}_{N|N}$.

Using either IRW or SRW models for the parameter variation, or more general versions of models I or II, similar behaviour is exhibited. However, the analysis is somewhat more complicated, and will not be pursued here.

Now bearing in mind the effect of using different values of W in the estimation, it is possible to employ a non-analytic procedure for choosing W. If we use the interpretation of Q as an *a priori* quantification of the rate of parametric change, then the diagonal elements have immediate meaning as the rates of individual parameter change. However, the off-diagonal elements are harder to interpret. Therefore one possibility would be to estimate the parameters $\underset{\sim}{\theta}_k$ with diagonal W, and using a number of different combinations of values of the diagonal elements. The results could then be examined, with the criteria for establishing the 'correct' value of W being largely based on the physical plausibility of the results obtained (Norton, 1975). Another possibility for choosing W would be to use $W = aS_N$, for various values of the scalar a, where S_N is as defined in Section 5.2.1. This has been the approach taken to estimating parametric change when using recursive IV methods (Young et al., 1971).

Although these procedures may appear somewhat *ad hoc* they provide a large amount of freedom for the experimenter to examine various hypotheses relating to the parameter movements, through the use of different values of W. At the same time, the results obtained in this way are subject to automatic constraints, so that it is not possible to obtain arbitrary estimates for the parameters. For example in the model (5.3.1) in which we are in effect estimating a time-varying mean of a series of observations, the range of possible trajectories for

$\{\hat{\theta}_k\}_{k=1}^{N}$ is between

$\hat{\theta}_{k|N} = \hat{\theta} = \bar{y}$ (with Q = 0, giving a constant mean)

and $\hat{\theta}_{k|N} = y_k$ (with $Q = \infty$, giving the mean at time k as y_k, the mean of a sample of size one at each time point)

It is also possible to develop analytic means for obtaining Q and σ^2 (or W). For this purpose, it is necessary to assume that the true parameter variation is of the form (2.3.2). If this is not the case, the same methods can still be employed, although their validity is largely diminshed. We first consider model I.

Under condition 1 of Chapter 3, that is, the Gaussian assumption on $\underset{\sim}{\theta}_0$; ν_k, e_k, k=1,2,...,N, the likelihood function for the sample $y_1, y_2, \ldots, y_N$ can be obtained as in Schweppe (1965). Following that author, we define $\lambda(k) = \log p(\underset{\sim}{Y}_k)$, the log-likelihood function. We can then put

$$2\lambda(k) = \log(2\pi)^k \det G_k - \underset{\sim}{y}_k{}^T G_k{}^{-1} \underset{\sim}{y}_k, \quad k=1,2,\ldots,N$$

Here G_k is the variance covariance matrix of $\underset{\sim}{Y}_k$. Now the joint density of $\underset{\sim}{Y}_k$ can be written as

$$p(\underset{\sim}{Y}_k) = p(\underset{\sim}{Y}_{k-1}) p(Y_k | \underset{\sim}{Y}_{k-1})$$

so that, taking logs, we obtain

$$\lambda(k) = \lambda(k-1) + \log p(Y_k | \underset{\sim}{Y}_{k-1})$$

The mean of the random variable $Y_k|\underset{\sim}{Y}_{k-1}$ is the conditional expectation $E(Y_k|\underset{\sim}{Y}_{k-1})$, so that we obtain

$$p(Y_k|\underset{\sim}{Y}_{k-1}) = \frac{1}{\sqrt{2\pi V_k}} \exp\left(- \frac{\varepsilon_k^{\,2}}{2V_k}\right)$$

where ε_k is the linear innovation first introduced in (2.3.3) and $V_k = V(\varepsilon_k)$, the innovations variance.

Thus

$$2(\lambda(k) - \lambda(k-1)) = - \log 2\pi V_k - \varepsilon_k{}^2/V_k$$

which finally yields

$$2\lambda(N) = - \sum_{k=1}^{N} \log 2\pi V_k - \sum_{k=1}^{N} \varepsilon_k{}^2/V_k \qquad (5.3.3)$$

This has, in fact, achieved a diagonalization of the quadratic form $\underset{\sim}{y}_N{}^T G_N{}^{-1} \underset{\sim}{y}_N$, through the use of the innovations process.

From Kailath and Frost (1968)

$$V_k = \underset{\sim}{u}_k{}^T P_{k|k-1} \underset{\sim}{u}_k + \sigma^2,$$

so that finally, we have

$$-\lambda(N) = \frac{1}{2} \{N \log 2\pi + \sum_{k=1}^{N} \log (\sigma^2 + \underset{\sim}{u}_k{}^T P_{k|k-1} \underset{\sim}{u}_k)$$

$$+ \sum_{k=1}^{N} \varepsilon_k /(\sigma^2 + \underset{\sim}{u}_k{}^T P_{k|k-1} \underset{\sim}{u}_k)\} \qquad (5.3.4)$$

Since the innovations ε_k and their variance V_k, for k=1,2,...,N can be obtained by successively estimating $\theta_{1|1}, \theta_{2|2}, \ldots, \theta_{N|N}$ with only a knowledge of W (Section 3.2), the log-likelihood

(5.3.3) can be expressed as a function of σ^2 and W. Then, to a constant,

$$\lambda(N) = -\frac{1}{2}\sum_{k=1}^{N} \log \sigma^2 T_k - \sum_{k=1}^{N} \varepsilon_k{}^2/\sigma^2 T_k \qquad (5.3.5)$$

where $T_k = V_k/\sigma^2$, an implicit function of W. Garbade (1977) considers the forms (5.3.5) of the log-likelihood, and setting $\frac{\partial\lambda(N)}{\partial(\sigma^2)} = 0$, obtains the concentrated log-likelihood function

$$\lambda^* = N \log \hat{\sigma} - \frac{1}{2}\sum_{k=1}^{N} \log T_k \qquad (5.3.6)$$

where $\hat{\sigma} = \left(\frac{1}{N} \sum_{k=1}^{N} \varepsilon_k{}^2/T_k\right)^{\frac{1}{2}}$

λ^* is now only a function of W, and in theory can be maximized with respect to this matrix, to obtain a maximum likelihood estimate for W. However, as Garbade points out, this is not a simple matter in practice. The difficulties are twofold : firstly, the severely non-linear occurrence of W in λ^*; and secondly, the requirement that the maximization of the likelihood take place over all symmetric, non-negative definite (n.n.d.) mxm matrices W. While the former problem can generally be overcome via numerical techniques, the latter cannot, except of course when m = 1. Thus, once again, we seek a smaller class from which to choose W.

The obvious choice is to restrict W to be a diagonal, n.n.d. matrix, as above in this section. Then a 'grid search' procedure, for example, may obtain, to sufficient accuracy, the values of $W_{i,i}$, $i = 1,2,\ldots,m$ which maximize λ^*. The data can then be processed using this value of W to provide the filtered and smoothed estimates of the parameters $\underset{\sim}{\theta}_k$.

For model II, the innovation representation (5.3.3) of the likelihood is not exact, nor are the innovations obtained from a filtering run using any of the methods discussed in Chapter 4. However, they may be used as an approximation, and the likelihood thus obtained once again maximized with respect to W.

Norton (1975) outlines an alternative method of choosing W, and, once again, the task is simplified by restricting to a diagonal W. For such W, the quantities d_k and f_k, $k=1,2,\ldots,N$ are calculated from

$$d_k = y_k - \underset{\sim}{u}_k{}^T \Phi \hat{\underset{\sim}{\theta}}_{k-1|N}$$

$$f_k = y_k - \underset{\sim}{u}_k{}^T \hat{\underset{\sim}{\theta}}_{k|N}$$

d_k and f_k may be thought of, as respectively, the smoothed innovations and the smoothed residuals.[†]

Finally, the sum of squares of smoothed innovations, and the sum of squares of smoothed residuals are calculated, and the

[†] Although Norton (1975) simply refers to innovations and residuals ('noise') so that there is some ambiguity, he has indicated in a personal communication (1978) that the smoothed versions of these quantities are used.

following statistics formed :

$$R_s = 1 - \left(\sum_{k=1}^{N} f_k^2\right)/\left(\sum_{k=1}^{N} d_k^2\right)$$

$$R_0 = 1 - \left(\sum_{k=1}^{N} d_k^2\right)/\left(\sum_{k=1}^{N} \hat{e}_k^2\right)$$

where $\hat{e}_k = y_k - \underset{\sim}{u}_k^T \underset{\sim}{\hat{\theta}}$, the residual obtained from a model with constant parameters.

Under the assumption of a random walk model, with the correct value of W used to estimate the parameters, R_s is a measure of the proportion of the error in the one-step ahead smoothed prediction that is due to parameter variation rather than observation or estimation error. R_0 indicates the proportion, of the prediction error in a constant parameter model, not accounted for by estimating the parameters as a random walk.

Now, as indicated for the model (5.3.1), the estimated noise free output $\underset{\sim}{u}_k \hat{\theta}_{k|N}$ will tend to follow the observed data exactly, in the limit as $W \to \infty$. Therefore $R_s \to 1$ as $W \to \infty$. R_0, on the other hand, may attain a maximum value with respect to W. Indeed, the behaviour with respect to W is determined by that of $\sum_{k=1}^{N} d_k^2$. Thus if the true value of W, say W_0, is greater than zero, values of W which are too small will tend to give larger prediction errors than the true value, because the parameter variation is not being allowed for. Conversely, values of W which are larger than W_0 will tend to alter the parameter estimate at time k-1 by combining noise effects with the parameters, so there will once again be large prediction errors. Hence we might reasonably expect a maximum in R_0. This possibility is not made clear by Norton, (1975), who recommends choosing W so that R_0 is as large as possible, with R_s 'below a specified limit'. He suggests that for small W, R_s is near zero; and then, at a certain point, as W is increased, R_s increases rapidly. It is this level which is taken as the 'specified limit' (Norton, 1978).

This procedure for obtaining estimates of W is obviously not rigorous, as was the case with the maximum likelihood estimation. There is, however, a relationship between the two procedures. Upon examination of (5.3.3), it can be seen that the likelihood is given by

$$L(Q,\sigma^2) = (2\pi) - \frac{N}{2} \prod_{k=1}^{N} V_k^{-\frac{1}{2}} \exp \{- \tfrac{1}{2} \sum_{k=1}^{N} \varepsilon_k{}^2/V_k\} \quad (5.3.7)$$

A 'generalized least squares' procedure for obtaining Q and σ^2 would be one where the exponent in (5.3.7) is maximized with respect to Q and σ^2, while an 'ordinary least squares procedure' would be one where the same quantity is maximized, under the assumption $V_k = 1$, $k = 1,2,\ldots,N$. Therefore Norton's approach in maximizing R_0 is approximately an ordinary least squares procedure.

The justification for such a simplification is well known in the case of constant $\underset{\sim}{\theta}_k$. However, simulation results indicate that the least squares approximation as used by Norton does not perform as well as the maximum likelihood estimate in the estimation of Q and σ^2 (see Section 6.1).

While maximizing R_0 can be interpreted as an approximation to the maximum likelihood procedure, the use of R_s is not so clearly defined. Norton's observation of a sharp rise in R_s at a certain value of W may be possible to corroborate analytically, although the analysis would presumably be quite difficult.

The above arguments would appear to indicate that the most satisfactory theoretical means of obtaining W is via the maximization of the concentrated log-likelihood with respect to a diagonal W. However, even with such a W, this maximization may not be easy, if there are a large number of parameters to be estimated. The likelihood is not necessarily unimodal, so that in a high-dimensional parameter space, numerical procedures may be computationally expensive, and may not even give the true maximum.

For the applications to real data where the methods of estimating time-varying parameters are to be used, the aim of the procedures is to examine parametric change. The exact size of the change may not be crucial, as long as it is detected. Therefore, in general, there may not be a need for very accurate estimation of W. As will be seen in Section 6.1 in the 1-parameter case, both R_0 and λ^* appear to exhibit sharp increases as functions of W. Although after a certain point, λ^* decreases sharply, while R_0 remains flat, values of W which give R_0 or λ^* in the upper part of this region of sharp increase should provide estimates of the parameters $\underset{\sim}{\theta}_k$ which do not overly exhibit spurious variation due to the effect of the observation noise e_k. Therefore, a reasonable procedure, which avoids some computational effort, for the estimation of W is to calculate $\sum_{k=1}^{N} d_k^2$ for each value of $W_{i,i}$, i=1,2,...,m; then increase each of these in turn until there is comparatively little change in $\sum_{k=1}^{N} d_k^2$, and use this final value of $W_{i,i}$ in estimation.

5.3.2 Initialization parameters

In Section 3.1, it was indicated that the algorithms could be initialized with virtually any value of the parameters, and a large initial estimation error covariance matrix. In most situations, the convergence to near the true parameter value during the filtering run is rapid, with accompanying decrease in the estimation error covariance matrix. There is, in theory, a small bias resulting from initialization in this way. However, it is insignificant, and may be neglected asymptotically. Nevertheless, in situations where the ratio of N, the sample size, to p, the number of parameters in the model, is small,such as in econometric models, it may be desirable to initialize the algorithms in a more specific manner. This can be accomplished either by a maximum likelihood procedure, or via a block initialization. We will consider each of these in turn.

The log-likelihood (5.3.3) can be considered as a function of $\underset{\sim}{\theta}_0$ and $P_{0|0}$. It can then be maximized with respect to these parameters, as was done for W and σ^2. The consistency result of Pagan (1978) mentioned in Section 2.3 provides a theoretical justification for this procedure. However, even if his method of proof extended to the random walk model, the procedure may be difficult to implement. The computational problems associated with maximum likelihood estimation of W and σ^2 alone would certainly be increased with the higher dimensions parameter space.

For model II, it may be possible to extend the parameter space still further, to include initial conditions on the model variables. Then the input, output and noise terms, with negative indices, which were all taken as zero in Sections 4.2 and 4.3, would be estimated as unknown parameters. This has been achieved for constant parameter time-series models (Newbold, 1974), and, in small samples, there is apparently some advantage to be gained. Once again, however, the added complexity would appear to counteract any benefits which might be obtained.

Block initialization may be used if the parameter variation is quite smooth. This involves estimating the parameters as constant from the first p samples. The remaining N-p samples are then processed as before, using the initial conditions $\hat{\underset{\sim}{\theta}}_p, P_p$, the estimate and error covariance matrix obtained from the first p samples.

It should be noted here that for the maximum likelihood estimation of W outlined in Section 5.3.2, it may be advisable to use a block estimate to initialize, and then calculate the likelihood for the remaining N-p samples only. This eliminates possible large deviations which may arise in early values of ε_k, the filtered innovations, in (5.3.3). Garbade (1977) suggests an alternative block initialization procedure to accomplish this task, while in Norton (1975), the use of smoothed innovations performs the same function.

5.4 Stability of the Estimation Procedure

It has been shown so far that, under certain assumptions on (2.3.1) and (2.3.2), we can obtain minimum variance linear unbiased estimates (approximate for model II) of the parameters $\underset{\sim}{\theta}_k$ in a random walk model. We have not, however, made any mention of the behaviour of the estimation procedures under consideration, for large sample size.

To investigate asymptotic properties, we once again turn to the state estimation literature. Jazwinski (1970) discusses sufficient conditions under which the estimation error covariance matrix $P_{k|k}$, for a linear state estimator $\hat{\underset{\sim}{x}}_{k|k}$, is uniformly bounded. These conditions are firstly, the positive definiteness of $P_{0|0}$, and secondly, the conditions of *uniform complete observability* (UCO) and *uniform complete controllability* (UCC). These latter are that the matrices $O(K,k-N_1)$, for some N_1, and $C(k,k-N_2)$ for some N_2 can be bounded above and below uniformly in k, where for the model (3.1.1)-(3.1.2),

$$O(k_1,k_0) = \sum_{t=k_0}^{k_1} (\Phi^{t-k_1})^T \underset{\sim}{u}_t \underset{\sim}{u}_t^T \Phi^{t-k_1}$$

$$C(k_1,k_0) = \sum_{t=k_0}^{k_1-1} \Phi^{k_1-t-1} Q(\Phi^{k_1-t-1})^T$$

(see Cooley and Wall, 1976).

Jazwinski then shows that under the same conditions, the linear system obtained from (3.2.1)-(3.2.3) is *uniformly asymptotically stable* : that is, rewriting the equation (3.2.1) in the form

$$\hat{\underset{\sim}{\theta}}_{k|k} = \Psi_k \underset{\sim}{\theta}_0 + \Delta_k y_k$$

we have $||\Psi|| \rightarrow 0$ exponentially. This property ensures that for bounded y_k, the filtered estimate $\hat{\underset{\sim}{\theta}}_{k|k}$ is also bounded. If, in (3.2.2)-(3.2.3) (the so-called *Ricatti equations*) $\underset{\sim}{u}_k$ were not dependent on k, it would be possible to obtain the asymptotic values of $P_{k|k}$ and $P_{k|k-1}$, by setting $P_{k|k} = P_{k-1|k-1} = R$, say, and $P_{k|k-1} = P_{k-1|k-2} = S$, say; and then solving (3.2.2) and (3.2.3) for R and S (Kailath and Ljung, 1976). The values R and S are the error covariance matrices for the corresponding steady state process. However, if $\underset{\sim}{u}_k$ is *not* constant, as is normally the case, there appears to be no proven result regarding the asymptotic values of $P_{k|k-1}$ and $P_{k|k}$. Indeed, it is likely that such a result would be difficult to obtain because of the complexity of the problem. Note that if Q = 0, then $P_{k|k} = P_{k|k-1}$, and $P_{k|k} \to 0$ as $k \to \infty$, providing

$$\sum_{k=1}^{N} \underset{\sim}{u}_k \underset{\sim}{u}_k^T \to \infty \qquad .$$

For the model (5.3.1), the Ricatti equations may then be solved, to yield $R = (-W + \sqrt{W^2+4W})/2$, $S = (W + \sqrt{W^2+4W})/2$. Thus as might be expected, both asymptotic values are monotonically increasing functions of W.

By analogy with these examples, it seems reasonable to expect that, under certain conditions, the matrices $P_{k|k}$, $P_{k|k-1}$ will exhibit some limiting behaviour. Certainly, results from simulations would suggest that such behaviour does occur in many cases (see Fig. 5.1).

Kailath and Aasnaes (1974) have demonstrated sufficient conditions for stability which are weaker than the UCO and UCC conditions. However, it should be noted that necessary and sufficient conditions have not been obtained by any author, as yet. Therefore, in practice it may be found that a system which fails to even satisfy these weaker conditions does not in fact lead to unstable estimation.

The following provides an example of instability which may occur. The system (3.1.1), (3.1.2) was simulated, with

$$\begin{aligned} &\text{(i)} \quad m = 2,\ N = 100 \\ &\text{(ii)} \quad u_k^{(1)} = 1,\ k=1,2,\ldots,N,\quad \sigma^2 = 1 \\ &\text{(iii)} \quad Q = \begin{pmatrix} 0.001 & 0 \\ 0 & 0.001 \end{pmatrix} \\ &\qquad\ \Gamma = I_2 \end{aligned} \tag{5.4.1}$$

$u_k^{(2)}$ was simulated

(a) as pseudo-random binary noise : that is, equally probable occurrences of -1 and 1;

(b) as linearly increasing : $u_k = k$, $k=1,2,\ldots,N$.

It was found that when the input $u_k^{(2)}$ was as in (a), $P_{k|k}$ decreased rapidly from its initial value of $10^6 I_2$. (see Fig. 5.1). However, for the input $u_k^{(2)}$ as in (b), $P_{k|k}$ increased steadily, with $P_{100|100} = \begin{pmatrix} 5 \times 10^9 & -5 \times 10^7 \\ -5 \times 10^7 & 5 \times 10^5 \end{pmatrix}$

The reason for this instability appears to be the violation of the UCO condition, when the input is as in (b).

For model II, it is not strictly possible to discuss stability in terms of the UCO and UCC conditions. However if we consider equation (4.4.1) as a linear observation equation, with x_k assumed known, then the UCO and UCC conditions can be written down. If the system is unstable (in the sense that the estimated model has zeroes of $\hat{A}(z^{-1}) = 0$ inside the unit circle) then x_k will become unbounded, and in this case the UCO condition will not be satisfied. Behaviour similar to that of the above example (5.4.1) may occur, causing the estimation to become unstable.

There is, in theory, a non-zero probability of such an instability occurring, since the parameters are assumed to be normally distributed at any time point. However, this will not necessarily cause problems. If the true parameter variation is non-stochastic, and if it is such that the output of the system remains bounded, then A_k is likely to be estimated such that the zeroes of $\hat{A}(z^{-1}) = 0$ lie outside the unit circle. Moreover, even if the true parameter variation is such that x_k eventually becomes unbounded while the noise level remains constant, improved parameter estimates may be obtained for a time because of the increased signal to noise ratio (see Lee, 1964).

Convergence of $P_{1,1}$ *in (5.4.1)(a)*

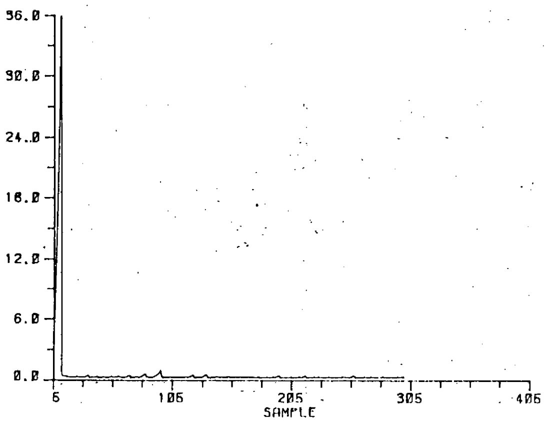


FIGURE 5.1

CHAPTER 6 : EXAMPLES

6.1 Simulation Results

In this section, some of the more important points discussed in earlier chapters will be exemplified using computer simulations. The presentation of the results will mostly take graphical form, as this appears to convey the relevant features most lucidly.

6.1.1 The random walk models for parameter variations

In order to consider the 'natural' properties of the three types of random walk models (RW, IRW, SRW) for parameter variation proposed in Section 2.2, each was simulated over 100 samples, with p=1. The same sequence ν_k, k=1,2,...,100 (as in (2.3.1)) was used in each simulation. Fig. 6.1 shows the resulting RW, Fig. 6.2 the IRW, and Figs. 6.3 and 6.4 show the SRW with, respectively, $\alpha = 0.9$, and $\alpha = 0.99$. It is clear that the RW exhibits 'jagged' variation, while the IRW appears to have a great deal of 'inertia' - once it is moving either up or down, it does not change direction, for a relatively long period. As $\alpha \rightarrow 1$, the SRW follows the IRW in shape, although of course not in dimension. The SRW with $\alpha = 0.9$ appears to be a useful model for tracking smooth parametric change. While it exhibits smooth variations, it does have the ability to change direction relatively quickly.

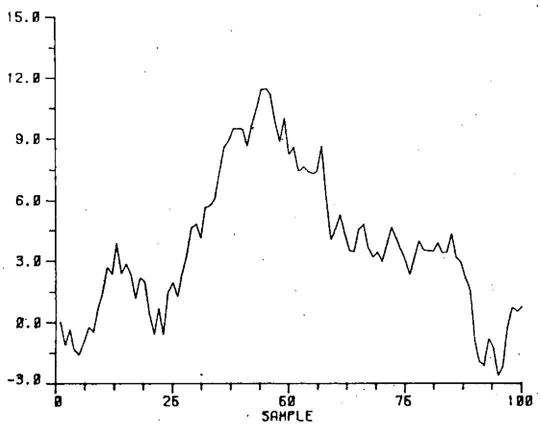


FIGURE 6.1

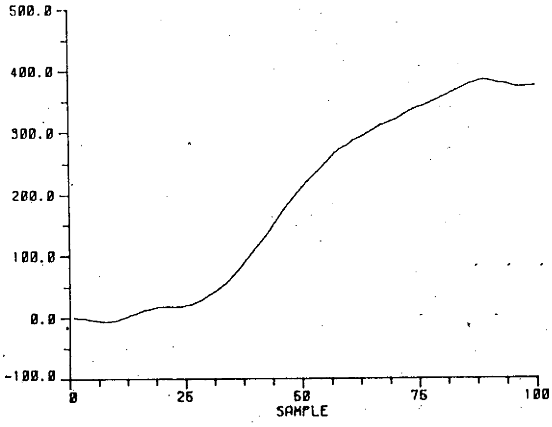


FIGURE 6.2

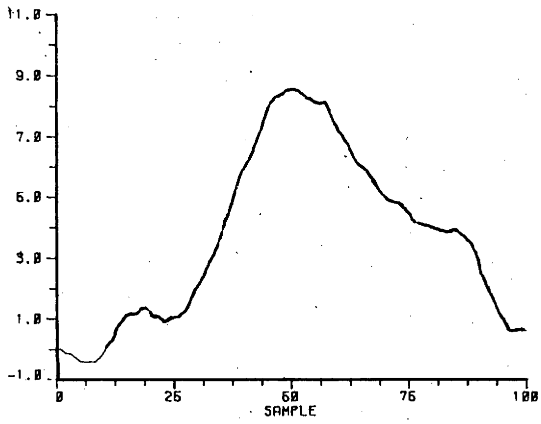


FIGURE 6.3

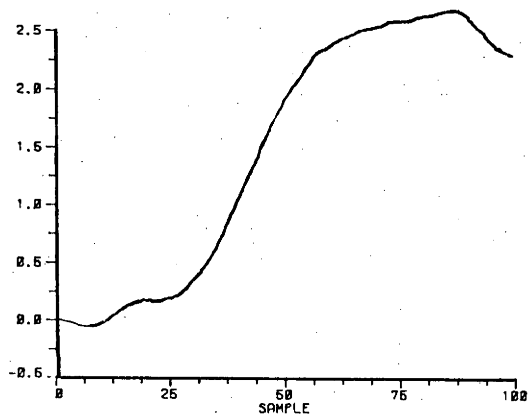


FIGURE 6.4

6.1.2 Instrumental variable estimation in model II

It would be possible to use many different examples to illustrate the filtering/smoothing algorithms of chapters 3 and 4. However, as a number of such simulations have been published (Lee, 1964; Young, 1969; Norton, 1975; 1976) we will restrict attention here to the results obtained from the new instrumental variable smoothing method of parameter tracking suggested in the present dissertation (Section 4.4).

The model chosen was as in (4.1.1), with

$$\frac{B(z^{-1})}{A(z^{-1})} = \frac{b_{0k}}{1 + a_{1k}z^{-1} + a_{2k}z^{-2}} \qquad (6.1.1)$$

This system was simulated over 100 samples, with σ^2 adjusted to give a signal to noise ratio of approximately 10 : 1. Initially, the true parametric variation was set as

$$b_{0k} = 0.15 + 0.05 \cos (\pi k/100)$$

$$a_{1k} = -0.4 + 0.05 \cos (\pi k/100)$$

$$a_{2k} = 0.5$$

$$k = 1,2,\ldots,100.$$

Both the IRW (Fig. 6.5) and the SRW (Fig. 6.6) were used to track the parameter variation. As can be seen, in both cases the least squares estimate (Section 4.2) provided a biased estimate of the variation in the parameter a_{1k}. However this estimate was largely improved by the use of the instrumental variable estimation. It was found that the iterative procedure mentioned in Section 4.4 had relatively little effect after the first iteration. Although parameter b_{0k} was also tracked very well, the results obtained are not presented, because the least squares estimate is not biased in this case.

Figure 6.7 shows a refined instrumental variable estimation of the same model (see Section 4.5) using the SRW. It can be seen that only a slight improvement over the least squares estimate (Fig.6.6) is obtained. The additional

Instrumental Variable Estimation - The parameter a_1 in (6.1.1)

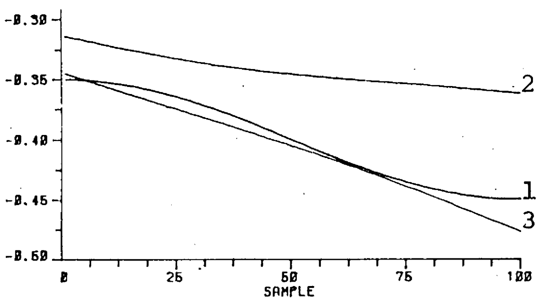


FIGURE 6.5

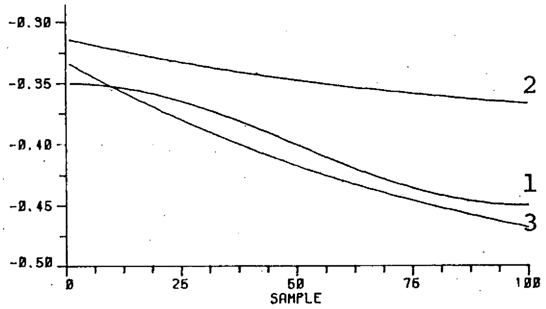


FIGURE 6.6

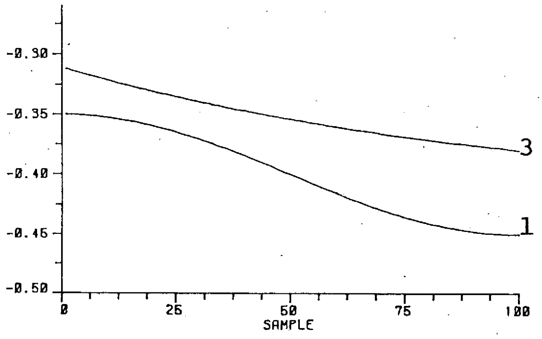


FIGURE 6.7

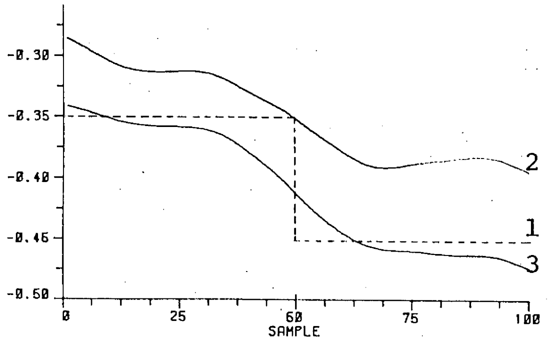


FIGURE 6.8

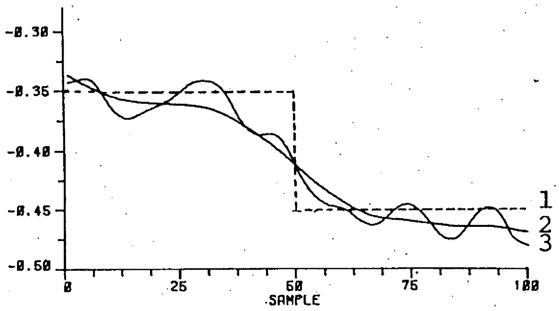


FIGURE 6.9

KEY FOR FIGURES 6.5 - 6.8

1 True parameter variation

2 Least squares estimate

3 Instrumental variable estimate

KEY FOR FIGURE 6.9

1 True parameter variation

2 SRW, $W_1 = 0.1$

3 SRW, $W_1 = 20.0$

In all cases W was taken as diag(W_1, W_2, W_3), with $W_i = 0$ for parameters assumed constant.

complexity of the refined form appeared to have a detrimental effect on the estimation, and it was found that the ordinary instrumental variable method was much more robust for general applications.

Figure 6.8 shows an instrumental variable estimation of the model (6.1.1) using an IRW, with now

$$b_{0k} = 0.15 \qquad k=1,2,..,100$$

$$a_{1k} = \begin{cases} -0.35 & k=1,2,...,50 \\ -0.45 & k=51,52,...,100 \end{cases}$$

$$a_{2k} = 0.5 \qquad k=1,2,..,100.$$

The signal to noise ratio was once again 10:1. The difficulties inherent in tracking the step change in a_{1k} can be clearly seen here. While there is definite evidence of such a change, it appears to have been 'smoothed' to a large extent. Once again, the bias in the least squares estimate is apparent. The same variation was also tracked with the SRW; Fig. 6.9 shows the result for two different levels of W_1. When $W_1 = 0.1$, the parameter is tracked too smoothly, as occurred for the IRW. For $W_1 = 20.0$, spurious variation is estimated due to noise effects, although the step appears more acutely.

The results shown in Figs. 6.5 - 6.9 are typical of those obtained from a number of simulations of parametric variation in model II. They usefully illustrate a number of the main features of the instrumental variable smoothing method of parameter tracking.

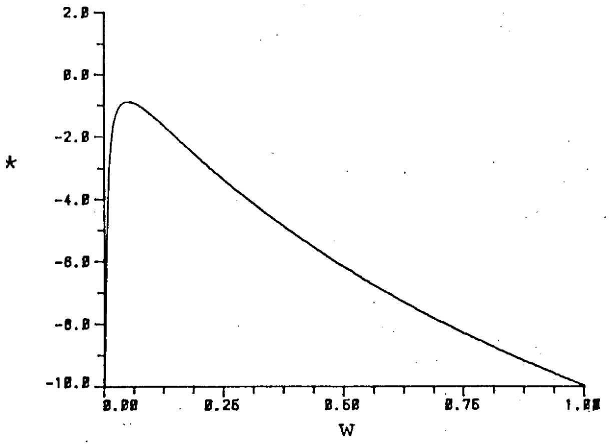


FIGURE 6.10

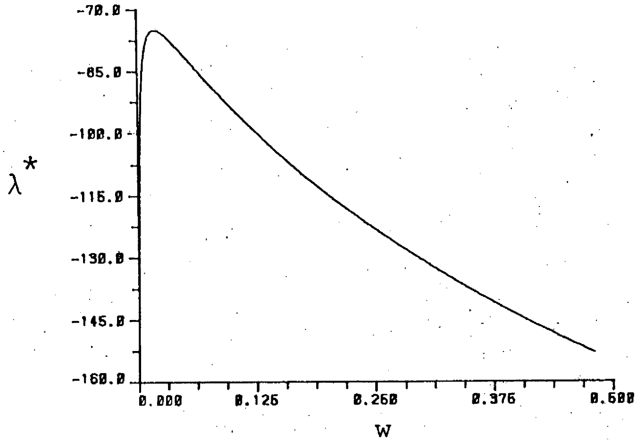


FIGURE 6.11

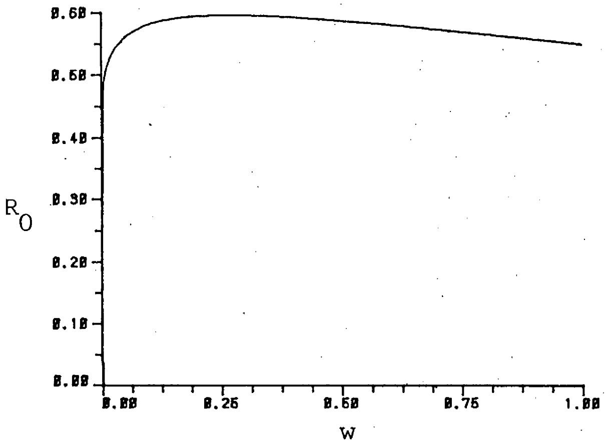


FIGURE 6.12

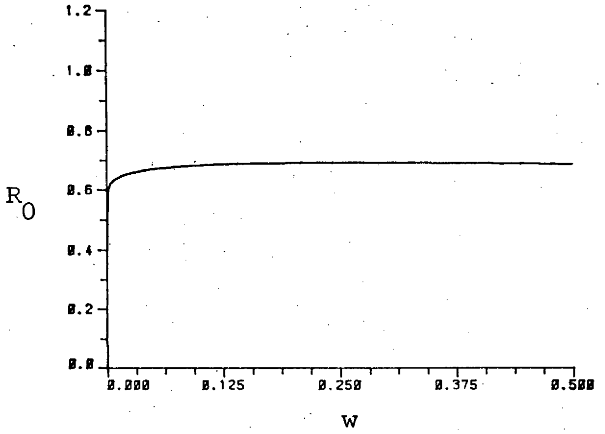


FIGURE 6.13

6.1.3. Estimation of W

The model (5.3.1) was simulated over 100 and 1000 samples. For convenience, the model (5.3.1) will be repeated:

$$y_k = \theta_k + e_k$$

$$\theta_k = \theta_{k-1} + \nu_k$$

$V(e_k) = 1$, $V(\nu_k) = 0.01$, $\theta_0 = 0$ were used here.

The full likelihood λ^* in (5.3.6), and the statistic R_0 of Norton (1975) were calculated, for a grid of values of W. As can be seen in Fig. 6.10, there is a distinct peak in λ^*, although the maximum likelihood estimate of W is somewhat biased. With 1000 samples (Fig. 6.11) the peak is even more distinct, and the bias has been reduced. On the other hand, Fig. 6.12 shows that R_0 attains a badly defined maximum. In the larger sample (Fig. 6.13) there appears to be very little improvement. Difficulties encountered with the maximum likelihood choice of W for real data will be illustrated in Section 6.2.

6.2 Analyses of Real Data

The range of possible applications of the methods discussed in this thesis is clearly very wide. Young (1969) has applied the techniques to the tracking of parameters in aerospace vehicle and chemical process models, and later (1974) in hydrological models. Norton (1975) has estimated time-varying response characteristics in a rainfall-runoff model. Finally, Garbade (1977) has used the procedures in an analysis of the demand for money in the United States. Some further simple analyses are presented here, with the accent on the use of the smoothing algorithms.

6.2.1 Rainfall trend analysis

There has been much discussion, in recent years, concerning the trends in rainfall patterns in south-eastern Australia. While

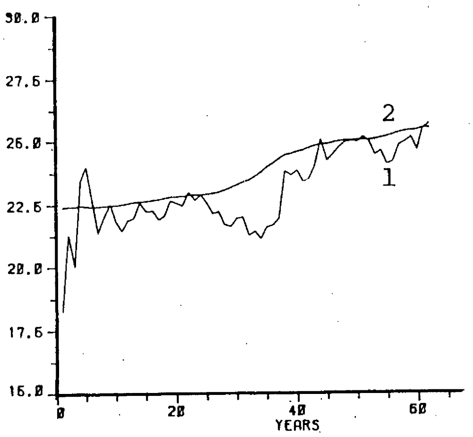


FIGURE 6.14

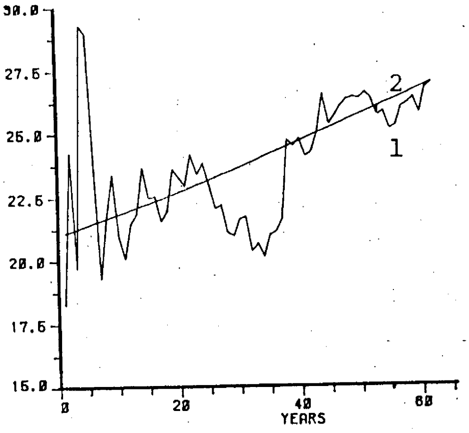
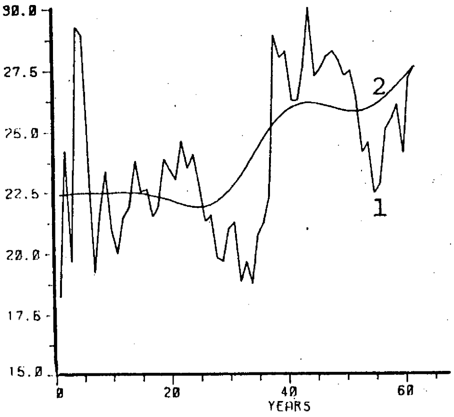


FIGURE 6.15



KEY

1 Filtered estimate

2 Smoothed estimate

FIGURE 6.16

some meteorologists (e.g. Pittock, 1975) suggest that a sharp increase in mean annual rainfall occurred in this area around 1945, conventional statistical testing (Gani, 1975) has tended to repudiate this theory.

In order to examine possible trends in the rainfall, annual records from a number of stations were examined. At each station, the annual rainfall was modelled as

$$y_k = \theta_k + e_k$$

where 0_k follows a random walk (2.3.2), and e_k is as in (1.1.1). The results obtained for Station 65 (Dubbo area) are typical of those obtained, and will be used to illustrate the analysis. The record available in this case was 62 years long, starting from 1913. The maximum likelihood method of Section 5.3 was used to estimate W, and the filtered and smoothed estimates of θ_k obtained using this choice of W are shown in Figs. 6.14 and 6.15 for, respectively, the RW and the IRW model of parameter variation. Fig. 6.14 shows a clear increase in the estimated (smoothed) mean around 1945. For the IRW, however, the maximum likelihood method appears to obtain a value of W which is too small : because the variation in the mean is apparently step-like, the 'average variation' over the whole sample is very small, so that the maximum likelihood estimate of W gives oversmoothing of the mean estimate. In fact, it appears that the increment S_k is estimated as constant, thus providing the result of Fig. 6.15. Figure 6.16 again shows the IRW estimates, this time with a much larger value of W chosen. Clearly, this result is more physically plausible, even though the result of Fig. 6.15 was obtained by the more rigorous maximum likelihood method. This demonstrates the dangers involved in placing too much faith in theoretically 'optimal' methods which may be restricted by the assumptions required in their development.

6.2.2 A simple air quality model

Half-hourly data on carbon monoxide concentration levels and wind speed were available for a station in the Canberra metropolitan area, and the simple model

$$y_k = \theta_k u_k + e_k$$

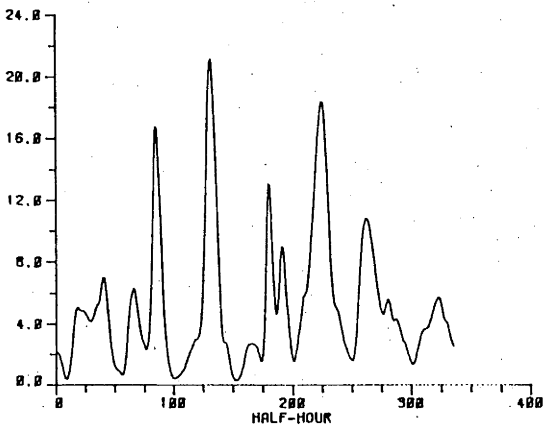
was proposed, where

y_k = carbon monoxide concentration in ppm

u_k = inverse of wind speed in m/sec.

Again, e_k is as in (1.1.1).

Estimating θ_k in this model as an IRW, using data for one week (starting 0000 hours, Monday) produced a smoothed estimate as in Fig. 6.17. Although no traffic flow data were available for the corresponding time period, it is apparent that the parameter is related to some variable of this kind. This suggests, as we would expect from physical principles, that an adequate model of carbon monoxide concentration would need to include traffic flow rate. Although in this case such a conclusion may be considered obvious, it is apparent that the concept can be used in many similar situations to ascertain relationships between variables, or to suggest whether data on additional variables should be collected (see Young, 1977).



Air Quality Model

FIGURE 6.17

CHAPTER 7 : CONCLUSION

In the preceding six chapters, we have systematically worked towards the development of a framework for the detection and estimation of parametric change in the transfer function time-series model. The regression model, which has been the object of most of the earlier work in this area, has provided methods which have then been extended for use with the transfer function model. Similarly, existing filtering algorithms for estimating parametric change in the transfer function model have guided the way to the development of the smoothing algorithms for this model. Subordinate to this primary aim has been the secondary objective of unifying a number of techniques - some analytically based, some *ad hoc* - which can be employed in the detection of parametric change.

There are a number of areas where future work could be carried out. In order to investigate parametric change in multivariable models, or models with coloured observation noise, the refined IV-AML procedure of Jakeman and Young (1978) could be adapted to incorporate a random walk model of parameter evolution. However, because of the increased complexity, it is doubtful whether useful results could be obtained in this framework. Rather, the simpler models discussed in this thesis could be used to suggest whether a meaningful multivariable or coloured noise model of the system under study could be obtained.

Another area of possible future interest is in the selection of the matrix W. As described in Section 5.3, the rigorous methods available have quite severe practical limitationsin a number of situations.

Finally, simple models with time-varying parameters may provide useful approximations to more complex, non-linear models. The dominant modes of behaviour may still occur in the simpler model, while avoiding difficulties associated with the more complex models.

REFERENCES

Aasnaes, H.B., and Kailath, T., 1973, "An innovations approach to least-squares estimation, - Part VII : Some applications of vector autoregressive-moving average models", *I.E.E.E. Trans. Autom. Control,* 18, 601-7.

Aström, K.J., and Bohlin, T., 1966, *The Theory of Self-Adaptive Control Systems* (New York : Plenum Press).

Bennett, R.J., 1976, "Non-stationary parameter estimation for small sample situations : a comparison of methods", *Int. J. Systems Sci.,* 7, 257-75.

Box, G.E.P., and Jenkins, G.M., 1970, *Time-series Analysis, Forecasting and Control* (San Francisco : Holden Day).

Box, G.E.P., and Tiao, G.C., 1975, "Intervention analysis with applications to economic and environmental problems", *J. Amer. Stat. Assoc.,* 70, 70-9.

Brown, R.L., Durbin, J., and Evans, J.M., 1975, "Techniques for testing the constancy of regression relationships over time, with comments", *J. Royal Stat. Soc., Ser. B,* 37, 149-92.

Bryson, A.E., and Ho, Y-C., 1969, *Applied Optimal Control* (Ginn).

Cooley, T.F., and Wall, K., 1976, "Identification theory for time-varying models", *Computer Research Paper No. 127,* N.B.E.R.

Cox, H., 1964, "On the estimation of state variables and parameters for noisy dynamic systems", *I.E.E.E. Trans. Autom. Control,* 9, 5-12.

Draper, N.R., and Smith, H., 1967, *Applied Regression Analysis* (New York : John Wiley).

Duncan, D.B., and Horn, S.D., 1972, "Linear dynamic recursive estimation from the viewpoint of regression analysis", *J. Amer. Stat. Assoc.,* 67, 815-21.

Gani, J., 1975, "The uses of statistics in climatological research", *Search,* 6, 504-8.

Garbade, K., 1977, "Two methods of examining the stability of regression coefficients", *J. Amer. Stat. Assoc.,* 72, 54-63.

Hannan, E.J., 1971, "The identification problem for multiple equation systems with moving average errors", *Econometrica,* 39, 751-65.

Harvey, A.C., and Phillips, G.D.A., 1976, "The maximum likelihood estimation of autoregressive-moving average models by Kalman filtering", *Q.S.S. Discussion Paper No. 38,* University of Kent, Canterbury.

Hildreth, C., and Houck, J.P., 1968, "Some estimators for a linear model with random coefficients", *J. Amer. Stat. Assoc.,* 63, 584-95.

Jakeman, A.J., and Young, P.C., 1978, "Refined instrumental variable methods of recursive time series analysis, Part II : multivariable systems", *C.R.E.S. Report No. AS/R13,* Australian National University.

Jazwinski, A.H., 1970, *Stochastic Processes and Filtering Theory.* (New York : Academic Press).

Kailath, T., 1968, "An innovations approach to least squares estimation - Part I : linear filtering in additive white noise", *I.E.E.E. Trans. Autom. Control,* 13, 655-60.

Kailath, T., and Aasnaes, H.B., 1974, "Initial condition robustness of linear least squares filtering algorithms", *I.E.E.E. Trans. Autom. Control*, 19, 393-7.

Kailath, T., and Ljung, L., 1976, "The asymptotic behaviour of constant-coefficient Ricatti differential equations", *I.E.E.E. Trans. Autom. Control*, 21, 385-8.

Kalman, R.E., 1960, "A new approach to linear filtering and prediction problems", *Trans. A.S.M.E. Ser. D*, 82, 35-44.

Kendall, M.G., and Stuart, A., 1961, *The Advanced Theory of Statistics* (London : Griffin).

Kopp, R.E., and Orford, R.G., 1963, "Linear regression applied to system identification for adaptive control systems", *A.I.A.A. J.*, 1, 2300-6.

Lee, R.C.K., 1964, *Optimal Estimation, Identification and Control* (M.I.T. Press).

Mehra, R.K., 1970, "On the identification of variances and adaptive Kalman filtering", *I.E.E.E. Trans. Autom. Control*, 15, 175-84.

Neethling, C.G., 1974, *Ph.D. Thesis*, University of Cambridge.

Neethling, C.G., and Young, P.C., 1974, "Comments on 'Identification of optimum filter steady state gain for systems with unknown noise covariances'", *I.E.E.E. Trans. Autom. Control*, 19, 623-5.

Newbold, P., 1974, "The exact likelihood function for a mixed autoregressive moving average process", *Biometrika*, 61, 423-6.

Norton, J.P., 1975, "Optimal smoothing in the identification of linear time-varying systems", *Proc. I.E.E.*, 122, 663-8.

Norton, J.P., 1976, "Identification of optimal smoothing using integrated random walks", *Proc. I.E.E.,* 123, 451-2.

Norton, J.P., 1978, *Personal communication.*

Pagan, A., 1978, "A unified approach to estimation and inference for stochastically varying coefficient regression models", *Discussion Paper No. 7814,* C.O.R.E., Université Catholique de Louvain.

Pittock, A.B., 1975, "Climatic change and the patterns of variation in Australian rainfall", *Search,* 6, 498-504.

Plackett, R.L., 1950, "Some theorems in least squares", *Biometrika,* 37, 149-57.

Rauch, H.E., Tung, F., and Striebel, C.T., 1965, "Maximum likelihood estimates of linear dynamic systems", *A.I.A.A. J.,* 3, 1445-50.

Rosenburg, B., 1972, "The estimation of stationary stochastic regression parameters re-examined", *J. Amer. Stat. Assoc.,* 67, 651-4.

Sage, A.P., and Ewing, W.S., 1970, "On filtering and smoothing algorithms for non-linear state estimation", *Int. J. Control,* 11, 1-18.

Sage, A.P., and Melsa, J.L., 1971[1], *Estimation Theory with Applications to Communication and Control* (McGraw-Hill).

Sage, A.P., and Melsa, J.L., 1971^2, *System Identification* (New York : Academic Press).

Salmon, M.H., 1978, "Recursive estimation as an aid to specification analysis in econometrics", *C.R.E.S. Working Paper No. R/WP29,* Australian National University.

Schweppe, F.C., 1965, "Evaluation of likelihood functions for Gaussian signals", *I.E.E.E. Trans. Inf. Theory,* 11, 61-70.

Sherman, S., 1955, "A theorem or convex sets with applications", *Ann. Math. Stat.,* 26, 763-7.

Soderström, T., Ljung, L., and Gustavsson, I., 1974, "A theoretical analysis of recursive identification methods", *Report No. 7427,* Lund Institute of Technology, Divison of Automatic Control.

Solo, V., 1978, "A unified approach to recursive parameter estimation", *C.R.E.S. Report No. AS/R20,* Australian National University.

Sorenson, H.W., and Stubberud, A.R., 1968, "Recursive filtering for systems with small but non-negligible non-linearities", *Int. J. Control,* 7, 271-80.

Swamy, P.A.V.B., 1971, *Statistical Inference in Random Coefficient Regression Models* (New York : Springer-Verlag).

Young, P.C., 1965,

Report No. PCY/TN(Camb)/1, Department of Engineering, University of Cambridge.

Young, P.C., 1969, *Ph.D. Thesis,* University of Cambridge.

Young, P.C., 1974, "Recursive approaches to time-series analysis", *Bull. Inst. Math. App.,* 10, 209-24.

Young, P.C., 1976 , "Some observations on instrumental variable methods of time-series analysis", *Int. J. Control*, 23, 593-612.

Young, P.C., 1977, "A general theory of modelling for badly defined systems", *C.R.E.S. Report No. AS/R9*, Australian National University.

Young, P.C., and Jakeman, A.J., 1978, "Refined instrumental variable methods of recursive time-series analysis, Part I : single-input, single-output systems", *C.R.E.S. Report No. AS/R12*, Australian National University.

Young, P.C., Jakeman, A.J., and McMurtrie, R.E., 1978, "An instrumental variable method for model structure identification", *C.R.E.S. Report No. AS/R22*, Australian National University.

Young, P.C., and Kaldor, J.M., 1978, "Recursive estimation as a tool for investigating climatic change", *C.R.E.S. Report No. AS/R14*, Australian National University.

Young, P.C., Shellswell, S.H., and Neethling, C.G., 1971, "A recursive approach to time-series analyses", *Report No. CUED/B - CONTROL/TR16*, Department of Engineering, University of Cambridge.