1. 12.75. Positive Microfilm Supplied to Keio University, Matsushita Memorial Library, Japan

A STUDY OF ESTIMATION PROCEDURES FOR TIME SERIES MODELS IN ECONOMICS

by A.R. Pagan

A thesis submitted for the degree of

Doctor of Philosophy in the Australian National University

The results presented in this thesis are my own except where otherwise stated.

G. R. Pagan (13/3/72)

Abstract

The thesis is concerned with the formulation and estimation of the autoregressive-moving average (ARMA) model, and its application to econometrics. Chapter 1 considers the origin of ARMA models and provides a discussion on the loss of optimal properties by a number of estimators under such a specification. Having established an à fortiori case for the ARMA model in economics, Chapter 2 derives the likelihood function in both the frequency and time domains, and outlines computational algorithms for its maximization - these being variants of the well known Gauss-Newton and Newton-Raphson techniques for the solution of systems of non-linear equations.

Chapters 3 and 4 contain Monte Carlo experiments on the estimators proposed in Chapter 2. These were primarily constructed to assess the likely impact of small samples upon the distribution of the estimators, but, as well, some indication of the sample size at which asymptotic theorems will hold is gained. As the first four chapters were concerned with single-equation problems Chapter 5 provides a generalization of the methodology to systems of equations, and reports on some experiments conducted with the systems estimator.

Finally Chapters 6 and 7 deal with applications of the ARMA model in economics. Chapter 6 demonstrates that a number of concepts appearing in economic theory e.g. permanent income, may be formulated as ARMA models, thereby enabling some estimates of these quantities to be made, while Chapter 7 compares the optimal ARMA estimator to ordinary least squares for a number of published studies, in order to demonstrate the extent of bias in conclusions based upon a use of the latter estimator. The final section of Chapter 7 outlines directions for future research.

ACKNOWLEDGEMENTS

My initiation into the subject of this thesis was performed by Professor A.W. Phillips and guidance into its mysteries by Professor R.D. Terrell. To both I owe a debt of gratitude. To the latter I especially wish to record my appreciation of his patient and helpful criticism throughout the preparation of the thesis.

Many others at A.N.U. and elsewhere should be thanked for their generous assistance. In particular I would like to mention the operators of the A.N.U. Computer Centre (Alan Harris, Errol Rumpf and Margaret Brinkley), Des Nicholls of the Department of Statistics (S.G.S.) and Ken Brewer and his colleagues at the Commonwealth Bureau of Census and Statistics. Finally I wish to express my appreciation of the painstaking efforts of Mrs I. Kewley in typing the manuscript.

Table of Contents

CHAPTER 2 The Estimation of ARMAX Models

Page

Tables

List of Figures

CHAPTER 1 The Origin of ARMAX Models

1.1 Introduction

An outstanding feature of econometric research has been the reliance upon the analysis of economic time series to yield valuable information about the parameters of economic models: a reliance occasioned by the inability of economists to perform experiments. The central feature of the time series utilized was the interdependence of observations in the available sample - an interdependence that queries the validity of applying many statistical theorems predicated upon the independence of observations. In order to take cognizance of this dependence (and in some instances to study the structure of the dependence itself) it was necessary to construct new theorems and tools of analysis. As they evolved they formed the nucleus of what is now referred to as time series analysis.

The influence of this corpus of techniques upon econometric specification and estimation appears to have waxed and waned repeatedly in the last half century. Some of the earliest studies of economic phenomena e.g. Beveridge [8] were applications of the tools of time series analysis and knowledge of this field seemed to be essential for econometricians. In the quarter century following World War II however, the interest of econometricians in this field seems to have died away (with the exception of the use of spectral analysis for the investigation of long swings and seasonal adjustment techniques) to the extent that until very recently few econometric textbooks mentioned

concepts such as correlogram, stationary process etc.[1]

What is the explanation of this curious state of affairs? Firstly it might be argued that the restrictive assumptions placed upon the evolutionary nature of the series did not seem to accord well with visual evidence on the behaviour of economic time series. Thus the notion of stationarity contained in (for example) Wold [119] did not seem apt for time series dominated by trend and seasonal movements. Secondly much of time series analysis was concerned with the analysis of a single time series and appropriate parametric descriptions of it whereas (perhaps stimulated by the Keynesian Revolution) economists wished to investigate the relationship between series. Lastly there were the difficulties raised in the estimation of the parameters of the specification considered apposite by time series theorists.

It will be the aim of this thesis to show that a knowledge of the concepts and tools of time series analysis is of importance to an econometrician and that with the aid of a digital computer these techniques may be easily integrated into a traditional econometric framework. To do this attention will be focussed upon the representation of time series as mixed autoregressive-moving average (ARMA) schemes. The constituents of this model- the autoregression and moving average parts - have a long history in time series analysis and Wold [119] is a

---

1 A glance at the early econometrics textbooks e.g. Tintner [106] and the standard reference books of the 'sixties Goldberger [30] and Johnston [56] emphasizes this point. One half of Tintner is concerned with the methods of time series analysis whilst the "moderns" have none. Of course the recent publication of Dhrymes [21] may rectify the imbalance.

convenient summary of many of the important applications. The symbolic form employed for an ARMA model will be

$$y(t) - \beta_1 y(t-1) - \ldots - \beta_p y(t-p) = \varepsilon(t) + \alpha_1 \varepsilon(t-1) + \ldots + \alpha_q \varepsilon(t-q) \qquad t=1,\ldots,N \qquad (1.1)$$

where $E(\varepsilon(t)) = 0$, $E(\varepsilon(t)\varepsilon(t-s)) = \sigma^2$ $s=0$; $= 0$ $s \neq 0$,

and E is the expectation operator.[2]

Equation (1.1) will be given the shorthand notation ARMA (p,q) where p is the length of the autoregression (A.R.) and q the length of the moving average (M.A.), and it describes the univariate, single equation formulation of an ARMA model.

Two further extensions of (1.1) are to be considered.

(a) Multivariate, single equation models of the type

$$y(t) - \beta_1 y(t-1) - \ldots - \beta_p y(t-p) = \gamma_o x(t) + \gamma_1 x(t-1) + \ldots + \gamma_r x(t-r) + \varepsilon(t) + \alpha_1 \varepsilon(t-1) + \ldots + \alpha_q \varepsilon(t-q) \qquad (1.2)$$

which is designated ARMAX (p,r,q) and the x(t) process has the property that

$$E(x(t)\varepsilon(t)) = 0$$

and may be regarded as an "exogenous" variable.[3]

---

[2] Henceforth any series designated as $\varepsilon(t)$ or e(t) will be assumed to possess these two properties i.e. zero expectation and finite variance. Sometimes the assumption will be made that the observations are identically and independently distributed (i.i.d.) otherwise it should be understood that any $\varepsilon(t)$ or e(t) is a normally and independently distributed (n.i.d.) random variable. Additional assumptions will be that y(t) is mean corrected i.e. $E(y(t)) = 0$, and that there is only one exogenous variable x(t). Both of these may be easily relaxed.

[3] The X in ARMAX refers to this point.

If there are a number of exogenous variables r will be a vector and the first and last subscripts in the brackets will be taken to refer to the order of the A.R. and M.A. respectively. An economic example of (1.2) might be to set (p=1, r=1, q=0) and to let y(t) be total consumer expenditure and x(t) total personal disposable income, thereby describing a consumption function frequently used in econometric models.

(b) Multivariate, Multiple Equation Systems

This will be restricted to allow interaction between the disturbances of each equation only, so that it represents a time series generalization of Zellner's "seemingly unrelated equations" estimator [123].

It will be argued in this and subsequent chapters that equations (1.1) and (1.2) (and the extension to multiple equation systems mentioned in (b) above) are appropriate to a wide variety of themes in modern econometric research.

A sketch of the chapter may now be given. Having established the type of model to be considered it is necessary to investigate

(i) The origin of these models i.e. are there any theoretical or empirical reasons for believing that equations (1.1) and (1.2) are likely to arise in the analysis of economic time series. Section 1.2 outlines various theorems that may be found in the time series literature which establish a presupposition in favour of ARMA models, and gives some illustrations of this from empirical research.

(ii) The conditions that must be placed upon the parameters of (1.1) and (1.2) and the evolutionary nature of the series y(t). It will be seen that the parameter restrictions are unusual and may

be quite strict. Section 1.3 then proceeds to examine the significance of the assumptions for the observed behaviour of a time series and asks if economic series are likely to be compatible with them?

(iii) Although there are a priori grounds for supporting the use of (1.1) and (1.2) we are led to enquire into the benefits that derive from an estimation of these equations. As economists are primarily interested in the magnitude of certain parameters, and the ability to predict certain variables, this section naturally concentrates upon the relationship of (1.1) and (1.2) to these problems.

(iv) Lastly we turn to the estimation of (1.1) and (1.2). In Section 1.5 the traditional tools of econometric analysis - Ordinary Least Squares and Generalized Least Squares - are analysed for their suitability for the estimation of the parameters of these equations. It will be seen (from numerical examples) that in a number of realistic models the losses (in terms of incorrect hypothesis testing) are very large from a use of Ordinary Least Squares and this fact prompts the development of estimators to avoid such losses.

1.2 The Origin of ARMA/ARMAX Models

1.2.1 Dynamic Behavioural Models

Static economics deals with instantaneous relationships between variables and summarizes such functions in time invariant measures such as elasticities and propensities. For a long while it was understood that there was not an instantaneous adjustment of one variable to

perturbations in another, but it was convincingly argued that measurements of parameters obtained under such assumptions would be approximately correct if the period of adjustment was shorter than the period at which the values of economic variables might be observed. Thus it was unlikely that many lags would be longer than a single year, which in general was the time period for which data was collected.

Even under this regime however there were some economic variables e.g.Investment in which the response would be distributed over a number of time periods, and it was to meet this objection that the concept of a distributed lag arose. With the ready availability of quarterly (and sometimes monthly) data this concept has become of crucial importance to applied econometricians as it is now evident that lags of the order of 10 quarters e.g. Investment may be required to adequately account for dynamic economic behaviour.

The general distributed lag specification is

$$y(t) = \sum_{j=0}^{\infty} w(j)\, x(t-j) + \varepsilon(t) \qquad (1.3).$$

A cursory inspection of (1.3) shows that the infinite number of parameters $w(j)$ prohibits the direct application of regression analysis. To escape from this quandry some writers recommended that $w(j)$ be set at zero for $j > K$ (where K is a finite integer set on a priori grounds) thereby constructing a finite distributed lag model which may be estimated by Ordinary Least Squares (O.L.S.). Unfortunately if K is at all large the presence of correlation in the $x(t)$ series will almost certainly lead to the cross product matrix being close to singularity.

Koyck [63] and Nerlove [80] had suggested that it was possible to transform an infinite series in w(j) into a finite A.R. and this prompted Jorgenson's rational lag function [57]. Defining the lag operator L with the property $L^m x(t) \rightarrow x(t-m)$ (1.3) may be rewritten as

$$y(t) = W(L)x(t) + \varepsilon(t) \qquad (1.4).$$

Then Jorgenson's theorems on rational lags state that if W(L) is an infinite (finite) polynomial in L an exact (arbitrarily close) representation of W(L) is given by

$$W(L) = \frac{C(L)}{B(L)} \qquad (1.5),$$

where $C(L) = \gamma_o + \gamma_1 L + \ldots + \gamma_r L^r$

$B(L) = 1 + \beta_1 L + \ldots + \beta_p L^p$

and r and p are small integers.

Using (1.5), (1.4) may be given the ARMAX (p,r,p) form

$$B(L)y(t) = C(L)x(t) + B(L)\varepsilon(t) \qquad (1.6).$$

From (1.6) it is apparent that ARMAX models arise naturally from certain methods of estimating distributed lag phenomena. Some popular versions of the polynomials C(L) and B(L) have been

(a) Koyck-Nerlove $\quad C(L) = \gamma_o \quad B(L) = 1 - \beta_1 L$

(b) Solow(Pascal) $\quad C(L) = \gamma_o \quad B(L) = (1 - \beta_1 L)^r$

where r is an integer in the range $1 < r << \infty$.

Both of these variants are well established in the literature but little attempt has been made to impose the ARMAX form: the most common explanation being that $B(L)\varepsilon(t)$ is white noise i.e. the initial

relationship contains correlated noise. Of course this assumption may be true but the form of the moving average required i.e. $B^{-1}(L)\varepsilon(t)$ is so restrictive as to make it unrealistic.[4]

Another behavioural relation that generates an ARMAX model is to be found in the literature centered on markets in which expectations (or anticipations) are important. The dearth of observations on such variables has made it customary to assume that expectations arise from past behaviour and may be replaced by a linear combination of past observable values.

To give an example, assume that there is a linear functional relation between a variable $y(t)$ and the anticipated value of a second variable $x(t)$. Denoting this anticipated value by $x^*(t)$ the function is

$$y(t) = \gamma\, x^*(t) + \varepsilon(t) \qquad (1.7).$$

To account for the formation of expectations concerning $x(t)$ from past values of $x(t)$ traditionally one of the following hypotheses has been adopted:-

(a) Adaptive Expectations: $x^*(t) - x^*(t-1) = \lambda(x^*(t) - x(t-1))$.

(b) Extrapolative Expectations: $x^*(t) = W(L)x(t)$.

The first of these was formulated initially by Cagan [16] in his study of hyperinflation and Nerlove in his study of supply response in the agricultural sector [80], while the second has been integrated into growth models by Nerlove and Arrow [84] and adopted by Turnovsky [108] in his investigation into the formulation of price expectations.

4 This does not rule out the possibility of a less restrictive M.A. in (1.4) so that the M.A. and A.R. parameters need not be equal in (1.6).

From the discussion of rational lags it is easily seen that (a) is a special case of (b) and that when substituted into (1.7) will give an ARMAX (1,1,1) model.

1.2.2 Dynamic Explanatory Models

The previous section has concentrated upon the origin of ARMAX models from behavioural relations. As well as this there is an argument from time series theory for these forms.

A beginning may be made by adopting Wold's decomposition theorem

"Summing up we have the following theorem in which one of the variables $\{\phi(t)\}$ and $\{\zeta(t)\}$ may be vanishing:

Theorem 7. Denoting by $\xi(t)$ an arbitrary discrete stationary process with finite dispersion, there exists a three-dimensional stationary process $\{\psi(t), \zeta(t), \eta(t)\}$ with the following properties

(A) $\{\xi(t)\} = \{\psi(t)\} + \{\zeta(t)\}$.

(B) $\{\psi(t)\}$ and $\{\zeta(t)\}$ are non-correlated.

(C) $\{\psi(t)\}$ is singular.[5]

(D) $\{\eta(t)\}$ is non-autocorrelated, and $E[\eta(t)] = E[\zeta(t)] = 0$.

(E) $\{\zeta(t)\} = \{\eta(t)\} + b_1\{\eta(t-1)\} + b_2\{\eta(t-2)\} + \ldots\ldots$

where b_n represents real numbers such that Σb_n^2 is convergent." [119,p.89]

For the present it is important to consider whether the deterministic component $\psi(t)$ vanishes for economic time series. One instance in which it would not is the presence of strict periodicities in the series. Nerlove has recently commented upon this in the following

[5] i.e. deterministic. A proof of the Wold theorem may be found in H. Cramer and M.R. Leadbetter. [19].

fashion:-

"For a long time seasonality was believed to be representable by a strictly periodic function,....However,...., such strict periodicity cannot plausibly be assumed to characterize economic time series" [82 p.12].

Additionally he argues in the following quotation that it is the linearly non-deterministic part which is of importance for the analysis of economic time series.

" We have found that most economic time series, possibly after some transformation to render them stationary, can be represented in the form (3.2)" (of the form (1.8) below).

Accepting the proposition that only the non-deterministic part is relevant, the series may be represented in the previous terminology as

$$y(t) = \sum_{j=0}^{\infty} w(j)\varepsilon(t-j) \qquad (1.8).$$

As with the infinite distributed lag it is not possible to estimate (1.8) directly, but it is possible to proceed by adopting the rational form of W(L) i.e. $W(L) = \frac{A(L)}{B(L)}$ so that (1.8) becomes the ARMA model[6]

$$B(L)y(t) = A(L)\varepsilon(t) \qquad (1.9).$$

It is worthwhile to note a special case of (1.9) i.e. A(L) = 1, $B(L) = 1 - \beta_1 L - \ldots - \beta_p L^p$ corresponds to the p'th order A.R. process.

---

[6] In this context the rational lag assumption is equivalent to the assumption that y(t) has a rational spectral density function - a not uncommon assumption in the fields of electrical engineering and acoustics e.g. see Yaglom [121].

This formulation has a long history in the analysis of stationary time series e.g. Kendall [60], Kendall and Stuart [61] and it is of some interest to enquire into the factors that might influence the choice of an ARMA rather than an A.R. model in parameterizing economic time series.

An obvious answer resides in the generality of the rational function employed in contradistinction to the fixed specification of $A(L) \equiv 1$ required by A.R. processes, and this generality should enable a wider class of series to be successfully modelled. A more pragmatic reason (and perhaps more appealing in view of the collinearity present between variables in many econometric models) is that an ARMA model may be the most parsimonious parameterization of the series. Box and Jenkins [9] have been the main proponents of this proposition and have justified such a claim by observing that there is an inverse relation between the order of the M.A. and A.R. required to yield the same representation of a series e.g. an infinite A.R. of the form $y(t) + \beta y(t-1) + \beta^2 y(t-2) + \ldots + \beta^k y(t-k) \ldots = e(t)$ may be converted to a first order M.A. $y(t) = (1-\beta L)e(t)$. Therefore low order ARMA models may be used in lieu of the high order A.R. processes that are generally required to characterize economic time series.[7]

From the previous discussion it is easily concluded that ARMA models are a natural functional form to reproduce a stationary time

---

[7] An example of this may be the fourth order A.R. used by Jorgenson, Hunter and Nadiri as a benchmark for the comparison of alternative econometric models of quarterly Investment behaviour [59]. As the data used was seasonally adjusted the fourth order A.R. term cannot be accounted for by the presence of a stable seasonal pattern.

series. However as most economists are aware few economic time series are stationary and most exhibit a trending mean and heteroskedasticity in the variance. To what extent then is the previous theory applicable? A partial answer comes from the widespread adoption of the "variate difference method" in the 1930's as a device for removing polynomial trends in the mean and variance of a series. Fishman [26 p.59] shows that for a series y(t) exhibiting a q'th order polynomial in the mean and variance an application of the filter $(1-L)^q$ will yield a stationary series with a constant mean and variance. The filtered series will be representable by an ARMA model. Although the "variate difference method" rests on assumptions that are untenable about economic series it is sufficient to note that differencing will reduce a non-stationary process to stationarity if repeated a sufficient number of times. Therefore by allowing B(L) to have roots on the unit circle (1.9) will be applicable to non-stationary series as well. More of this in Chapter 6 when we come to consider some actual time series.[8]

A second explanation of the source of ARMA models lies in what we will designate as time transform effects. Quenouille [94] was the first to isolate these by demonstrating that a continuous variable, when observed at intervals of p time units apart, would exhibit a correlogram identical to that of an ARMA (p,p-1) equation. Quenouille's

---

[8] Authors who have used differencing to induce stationarity are Yaglom [121], Whittle [117], Box and Tiao [12], Nerlove, Grether and Couts [18], Nerlove and Wage [83], Theil and Wage [104] and Box and Jenkins [9]. Of these Yaglom seems to have been the first. Box and Jenkins have the most discussion on the philosophy underlying it.

theorem is of some importance to econometrics as it demonstrates that ARMA models are the discrete analogues of relationships believed to hold in continuous time and suggest that they are a natural model to fit when only discrete observations are available.

Another view of this effect is gained by considering the form of A.R. processes under various time regimes. To give a simple example assume that y(t) follows a second order A.R.

$$y(t) + b_1 y(t-1) + b_2 y(t-2) = e(t) \qquad t = 1,\ldots,N \qquad (1.10).$$

Consider the change in (1.10) when information can be obtained for only every second time point i.e. we observe r = 2,4,..,N (if N is even). Even though y(t-1) is no longer observable it is possible to reformulate (1.10) as an A.R. in observable values only by the following method. Lag (1.10) once to get (1.11a) and lag (1.11a) once to arrive at (1.11b).

$$y(t-1) + b_1 y(t-2) + b_2 y(t-3) = e(t-1) \qquad (1.11a)$$

$$y(t-2) + b_1 y(t-3) + b_2 y(t-4) = e(t-2) \qquad (1.11b)$$

Solving (1.10), (1.11a) and (1.11b) the reduced form is

$$y(t) + (2b_2 - b_1^2)y(t-2) + b_2^2 y(t-4) = e(t) - b_1 e(t-1) + b_2 e(t-2) \qquad (1.12).$$

Re-indexing equation (1.12) in the r-domain we have

$$y(r) + (2b_2 - b_1^2)y(r-1) + b_2^2 y(r-2) = u(r) \qquad (1.13).$$

What is the autocorrelation pattern of u(r) in (1.13)? The expressions for u(r), u(r-1) and u(r-2) are

$$\begin{aligned} u(r) &= e(t) - b_1 e(t-1) + b_2 e(t-2) \\ u(r-1) &= e(t-2) - b_1 e(t-3) + b_2 e(t-4) \\ u(r-2) &= e(t-4) - b_1 e(t-5) + b_2 e(t-6) \end{aligned} \qquad (1.14).$$

From (1.14) the covariance function of u(r), Γ_{uu} is given by [9]

$$\begin{aligned} \gamma_{uu}(o) &= \sigma^2(1+b_1^2+b_2^2) \\ \gamma_{uu}(1) &= b_2\sigma^2 \\ \gamma_{uu}(j) &= 0 \qquad\qquad j = 2,\ldots,\infty \end{aligned} \tag{1.15},$$

so that since u(r) has the correlation function of a first order M.A. process (1.13) becomes [10]

$$y(r) + (2b_2-b_1^2)y(r-1) + b_2^2y(r-2) = \varepsilon(r) + \alpha_1\varepsilon(r-1) \tag{1.16}.$$

This simple example illustrates the transition between A.R. and ARMA forms as the period of observation changes. In general (see Appendix 1) it may be shown that for an A.R. of order p indexed on t = 1,...,N, the appropriate process indexed on r = k, 2k,...,N (if N is a multiple of k) is ARMA $(p, I(\frac{p(k-1)}{k}))$ where I is the operator truncating the expression in brackets to the integer immediately below this value.

A number of interesting cases emerge. Firstly if p = 2, k = 2 (the example given above) we obtain ARMA (2,1) as was found. Secondly as $k\to\infty$ we obtain the "balanced" ARMA (p,p) model. Lastly as $k\to 1$ i.e. the r and t domains coincide, the ARMA (p,o) model emerges i.e. the original A.R.

---

[9] In future chapters the covariance function of a series x(t) will be denoted by Γ_{xx} and the individual covariances by $\gamma_{xx}(j)$. The cross covariance between x(t) and y(t) will be Γ_{xy} with elements $\gamma_{xy}(j)$.

[10] See Box and Jenkins [9 p.68] for a discussion on the theoretical autocorrelation function of a M.A. process. A number of properties of autocorrelation functions will be assumed in the thesis as a good account is contained in Chapters 2 and 3 of that book.

Brewer [15] has extended the above analysis to a large number of different models. He shows that an initial ARMA (p,q) equation will become ARMA $(p, I(\frac{p(k-1)+q}{k}))$ if the variable observed is a stock and ARMA $[p, I(\frac{(p+1)(k-1)+q}{k})]$ if it is a flow. The reader is referred to his paper for an extensive discussion of the possible variations.[11] The importance of his contribution lies in the limits as $k \to \infty$ i.e. ARMA (p,p) (for stock models) and ARMA (p,p+1) (for flow models) signifying that <u>if ARMA models are to be justified by time transform effects alone the order of the M.A. can never be greater than the order of the A.R. for stock variables and only one greater for flow variables.</u> Such a principle may help to explain the heavy emphasis upon "balanced" models in the work of Box and Jenkins.

A source of high order autocorrelation that may require a complex disturbance assumption is that originating from the seasonal patterns present in many monthly and quarterly data series. If the error term is to reflect unobservable (or omitted) variables then it is likely that these patterns will be retained in the disturbances. Additionally the practice of incorporating dummy variables into regression equations will only serve to remove a <u>stable</u> seasonal pattern so that any evolutionary behaviour in this component will be reflected in the error term.

---

[11] Some mention should be made of the studies of Zellner [125], Zellner and Montmarquette [127], and Telser [103] who also arrive at ARMA or ARMAX forms by time aggregation. However in the case of the first two papers there is no need to estimate M.A. parameters because k is assumed known and there are no lagged dependent variables, so that the Aitken estimator based on the known covariance matrix will be fully efficient.

Thomas and Wallis [105] have considered this aspect and have concluded that for a quarterly model it is advantageous to test for the presence of fourth order serial correlation. To this end they hypothesize the disturbance specification

$$u(t)\ (1-\rho\ L^4) \ = \ e(t) \qquad (1.17)$$

which yields $\rho = \gamma_{uu}(4)/\gamma_{uu}(0)$. Their test statistic is basically a test that the estimated value of $\rho(=\hat{\rho})$ is statistically insignificant from zero and is a generalization of the Durbin-Watson d statistic to a higher order disturbance format. Of course the test will lack power as even the traditional specification

$$u(t)\ (1-\tilde{\rho}\ L) \ = e(t) \qquad (1.18)$$

has $\gamma_{uu}(4)/\gamma_{uu}(0) = \ (\tilde{\rho})^4$, so that it is possible to obtain a high fourth serial correlation coefficient from a model with a disturbance format not generated by (1.17).

These criticisms should not detract from Thomas and Wallis' contribution as it is becoming clear that as the data is observed at shorter intervals the autocorrelation patterns in the disturbances are potentially more complex.[12] However there must be some doubt about the adequacy of (1.17) as an appropriate description (say) of an omitted variable with a quarterly seasonal pattern. Allowance for the presence of trend factors in such series leads to the contention that a more plausible version would be

$$u(t)(1-\rho_1 L-\rho_4 L^4) \ = \ e(t) \qquad (1.19).$$

An even more fundamental criticism lies in the use of the autoregressive

---

[12] An example is provided by Trivedi's report on inventory equations [107] in which a M.A. of third order was normally required.

format of (1.17), (1.18) and (1.19) to account for autocorrelation. For example, assuming a regression model

$$y(t)(1-\beta L) = \gamma\, x(t) + u(t) \qquad (1.20)$$

where u(t) has the form (1.19), it is possible to rewrite (1.20) as

$$y(t)(1-\beta L)(1-\rho_1 L-\rho_4 L^4) = \gamma(1-\rho_1 L-\rho_4 L^4)x(t) + e(t) \qquad (1.21).$$

O.L.S. may be applied to (1.21) to yield consistent estimates of $\beta,\gamma,\rho_1,\rho_4$ as the disturbance term is no longer correlated with any of the regressors. Of course this has been achieved by adding y(t-2), y(t-4), y(t-5), x(t-1) and x(t-4) to the regression thereby reducing the disturbance term to white noise. In contrast to this it is impossible to remove a M.A. in this way, so that if consistent estimates are desired it would always be necessary to check that this type of autocorrelation was not present.[13]

1.3 Some Theorems on Time Series

Throughout this thesis attention will be focussed upon the estimation of the time series models discussed in the preceding two sections. The estimators devised for this purpose must possess certain properties if they are to be of use in econometric investigation and it is assumed that minimal requirements would be consistency and asymptotic normality. However as will become apparent later it is not easy to provide such proofs, and existing theorems are subject to a number of conditions pertaining to the evolutionary nature of time series and

---

[13] One suspects that in many econometric studies lagged dependent variables are being added because of this "autocorrelation reducing" feature rather than as a result of some postulated behavioural phenomenon.

the magnitude of the parameters of ARMA/ARMAX models. With reference to the former the most general conditions under which the properties of time series estimators seem to hold are Grenander's conditions (see Hannan [47 p.122]).

CONDITION 1 $\lim_{N\to\infty} d_j^2(N) = \lim_{N\to\infty} \sum_{t=0}^{N} x_j^2(t) = \infty \quad j = 1,\ldots,k$

CONDITION 2 $\lim_{N\to\infty} x_j^2(N) \,/\, d_j^2(N) = 0 \qquad j = 1,\ldots,k$

CONDITION 3 $\lim_{N\to\infty} \dfrac{\sum_{t=1}^{N} x_j(t)\, x_i(t+h)}{d_j(N)\, d_i(N)} = \rho_{ji}(h)$ exists $\quad i,j = 1,\ldots,k.$

and $\rho_{ji}(h)$ is the h'th population cross correlation coefficient between $x_j(t)$ and $x_i(t+h)$.

A wide variety of evolutionary types are encompassed in Conditions 1-3. It is asserted that most economic time series are either described by polynomial functions, or may be when suitably transformed. To support the first part of the assertion we refer to Tintner's demonstration that the variate difference method (which eliminates functions only up to and involving a polynomial) reduced many series to white noise, and the satisfactory performance of polynomials in regression models to account for trend and seasonal components (Henshaw [50]). For the second part it is sufficient to note that the only other evolutionary pattern of general interest for economic time series - the exponential - may be reduced to a polynomial by a logarithmic transformation of the data.

Nevertheless the assertion is not meant to be universal. Undoubtedly there are a number of economic series which cannot be transformed in a simple manner e.g. the demand for a new product frequently

follows a path involving exponential growth during only part of the product cycle, and it would be unwise to model these with the techniques of this thesis. Certainly a visual inspection of the graphed data is always recommended before any models are fitted.

Having decided on polynomial functions as suitable descriptions of economic series it is necessary to adopt a particular form for the ARMA/ ARMAX model in order to assess compatibility with Grenander's conditions. For this purpose unit roots in B(L) will be advocated. Choosing the simplest ARMA form $y(t)\ (1-L) = e(t)$, the covariances may be derived as

$$\gamma_{yy}(o) = N\sigma^2$$

$$\gamma_{yy}(j) = (N-j)\sigma^2.$$

Conditions 1 and 2 require the absolute magnitude of the observations to become larger with time (Condition 1) and that the latest observation must not dominate the series (Condition 2) and are clearly satisfied for any function up to an exponential. However Condition 3 requires an examination of the above correlations. In the example $\rho_{yy}(h) = 1 - \frac{1}{N}$, $\underset{N\to\infty}{\text{Lim}}\ \rho_{yy}(h) = 1$ and this limit will be approached from below thereby satisfying the third condition.

From these arguments we conclude that series which may be modelled adequately by including unit roots in B(L) will obey Grenander's conditions and most theorems of time series analysis will be applicable. As a test of whether polynomials are sufficient one may form the differenced series and investigate its behaviour : if there is an exponential term in the original series the filtered series should be non-stationary. For future reference it should be remembered that the

correlogram of a series capable of being represented by unit roots in B(L) will die out only very slowly (as seen from $\rho_{xx}(j) = 1 - {}^{j}/T$).

The above discussion intimates that the number of evolving series which do not satisfy Grenander's conditions is likely to be small and certainly the series to be studied in this thesis (primarily consumption) do not. A further discourse on this will be given in the chapters concerned with empirical modelling.

As well as the conditions imposed upon the nature of the time series the range of the parameters of ARMA/ARMAX models must be restricted, partly to achieve identification - a logical prelude to estimation. In the form given here the conditions are to be found in Hannan [46].

Let the ARMAX (p,r,q) model be

$$B(L)y = C(L)x + A(L)e$$

with B(L) of order p, C(L) of order r and A(L) of order q. Then four conditions must be imposed upon the polynomials B(L), C(L) and A(L).

CONDITION 4. There must be no roots common to B(L), C(L) and A(L). Some examples of the flat likelihoods (indicative of a lack of identification) associated with a violation of this assumption are to be found in Box and Jenkins [9], and these authors warn that in practice the existence of roots which are approximately equal will result in the same effect. An important model in which this occurs will be encountered in Chapter 6.

CONDITION 5. If the polynomials B(L), C(L) and A(L) are of order p,r and q respectively then β_p, γ_r and α_q cannot all be zero. This can be deduced from Hannan's general theorem [46] which when formulated for

non-simultaneous systems of equations requires that if the maximum orders in the system of G equations are p,r and q and we define G dimensional square matrices B(p), C(r), A(q) as those containing the β_p, γ_r, α_q parameters of each equation, then there exists no vector ϕ (not equal to a vector of zeroes) that has the property

$$\phi^T B(p) = \phi^T C(r) = \phi^T A(q) = 0.$$

Specializing to a single equation it is clear that ϕ is scalar and $\phi = 1$ will satisfy these relations simultaneously whenever $\beta_p = \gamma_r = \alpha_q = 0$. Furthermore a little thought will confirm that if one of these is non-zero only $\phi = 0$ will suffice. In practice though the existence of approximate zeroes is almost certain to result in a lack of identification.

CONDITION 6. The polynomial C(L) must have at least one non-zero element (Nicholls [85]). Consider a rational lag model. If C(L) = 0 there will be common roots in the M.A. and A.R. transfer functions and this will cause a lack of identification. The origin of such a difficulty could well be the erroneous inclusion of a particular x(t) into the model to be estimated.

CONDITION 7. The roots of the polynomial equations B(L) = 0, C(L) = 0 and A(L) = 0 must lie on or outside of the unit circle. Especially we shall require the roots of B(L) = 0 to lie either on or outside of the unit circle and of C(L) = 0 and A(L) = 0 to lie strictly outside the unit circle.

Condition 7 is not required for identification. To justify the restriction on B(L) we note that Grenander's conditions 2 and 3 stipulate

the admissible class of functions as being bounded from above by the exponential and that polynomial functions were associated with unit roots in B(L). If roots inside the unit circle were allowed it would be possible for Condition 2 (and almost certainly Condition 3) to be violated. Again we note that the presence of unit roots in B(L) seems to generate evolutionary moments that are in accord with those observed for economic time series. A strict condition has been imposed upon C(L) as the likely presence of non-stationarity (in the polynomial sense) in x(t) variables will then be preserved rather than accentuated.

Lastly the restriction on A(L) must be justified. The symmetric nature of covariance functions makes it impossible to distinguish between the correlogram of a M.A. with roots α and another with roots α^{-1} so that to achieve identification it is necessary to choose one of the two sets. A number of reasons may be given for the selection of the set that lies outside the unit circle.

(a) Box and Jenkins argue that it is necessary for the process to be invertible i.e. that a M.A. have an infinite A.R. representation, and the Laurent series expansion to achieve this will not exist unless all roots of A(L) = 0 are outside the unit circle.

(b) As will be seen later (section 2.6) the properties of estimators of ARMA equations have only been proven under the assumption that the equation can be written as a linear process <u>whose frequency response function possesses an inverse</u>. This is not the case for unit roots in A(L). It is also true that a unit root in B(L) will contradict this assumption but we may always remove these by differencing whereas this is not so for A(L).

(c) There is a difficulty in interpreting the significance of a unit root in A(L) for the behaviour of the series whereas for B(L) an identification with polynomial trends is available. Furthermore if B(L) has a unit root an attempt to estimate a unit root in both transfer functions would lead to a lack of identification in ARMA models.

Conditions 1-3 will be assumed to hold for all time series considered in this thesis and all parameters of ARMAX models will be constrained to satisfy Conditions 4-7.

1.4 Reasons for the Estimation of ARMAX Models

Section 1.2 has discussed the origin of ARMA/ARMAX models whilst Section 1.3 has outlined certain features that the time series and ARMA/ARMAX model selected must possess. In this section we concentrate upon the benefits to be gained from the estimation of the parameters of such models - a prelude to a more intensive examination of certain estimation procedures.

Two benefits will be claimed for ARMAX models[14]

(i) "Optimal" estimation.

(ii) "Optimal" prediction.

1.4.1. "Optimal"Estimation

A comprehensive account of this is given in Nerlove [81]. It is argued there that the generalized least squares (G.L.S.) and maximum

---

[14] We mean by this that all parameters of B(L), A(L) and C(L) are estimated in contrast to methods (such as O.L.S.) which would obtain only those of B(L) and C(L). As the latter will frequently be computationally simpler we seek the benefits of estimating A(L).

likelihood (M.L.) estimators are identical if $\varepsilon(t)$ in (1.1) and (1.2) are n.i.d. As such the G.L.S. estimator of the coefficient vector $\theta = (\beta,\gamma)$ (where β and γ are vectors of A.R. and exogenous variable parameters) is consistent and $\sqrt{N}(\hat{\theta}-\theta)$ is asymptotically normally distributed with mean zero and covariance matrix - $[E(\frac{\partial^2 L}{\partial\theta\partial\theta^T})]^{-1}$ where L indicates the likelihood and T signifies transposition.

Denoting the covariance matrix of $u(t)(=A(L)\varepsilon(t))$ by Ω the G.L.S. estimator maximizes the quadratic form $L = -u^T\Omega^{-1}u$. As Ω is a function of the M.A. parameters an estimate of these is important if estimates of $(\beta_1,\ldots,\beta_p)$, $(\gamma_1,\ldots,\gamma_r)$ are to be acquired (the type of estimate of α required i.e. consistent or efficient will be considered later).

A further advantage of an efficient estimator is the derivation of the correct covariance matrix. This is of crucial importance for hypothesis testing as the principal statistics e.g. the t-test involve the standard deviations of the estimated parameters, and econometricians have recognized for a considerable number of years (see Johnston [56]) that the presence of autocorrelation in the error term results in an estimate of the covariance matrix that may diverge substantially from the correct value.

1.4.2. "Optimal" Prediction

The theory of optimal signal extraction and prediction is presented in Whittle [117] and a summary for economists is contained in the article by Grether and Nerlove [36]. A statement of the principal theorem from this literature follows.

Let $y(t)$ be a stationary non-deterministic process with the M.A. representation

$$y(t) = \sum_{k=0}^{\infty} w_k \varepsilon(t-k) = W(L)\varepsilon(t) \qquad (1.22).$$

Then the minimum mean-square error linear prediction of $y_{t+\nu}$ at time t has the analytical expression:

$$y_{t+\nu,t} = \psi_\nu(L)y(t)$$

$$\psi_\nu(L) = \frac{1}{W(L)}\left[\frac{W(L)}{L^\nu}\right]_+$$

where $\left[\ \right]_+$ connotes an operator which retains only the non-negative powers of L in the bracketed expression. For an ARMA model $W(L) = \frac{A(L)}{B(L)}$ so that the prediction formula becomes

$$\psi_\nu(L) = \frac{B(L)}{A(L)}\left[\frac{A(L)}{B(L)L^\nu}\right]_+ \qquad (1.23).$$

Equation (1.23) emphasizes the dependence of the optimal predictor upon the M.A. parameters i.e. serial correlation of the errors, the implication being that such knowledge will lead to improved forecasts. An alternative (and more familiar) expression of this would be:

> "For the sample gives information, although imperfectly, about the values of the errors during the period of observation. Because of the serial correlation, it also provides some indication on subsequent values of the errors" [71 p.535] .

1.5 General Considerations Relating to the Estimation of ARMAX Models

In preceding sections analyses of the origin of the ARMAX model, the conditions to be placed upon it, and the be efits that would accrue from the estimation of the parameters associated with it have been

presented, and it is now appropriate to investigate a number of alternative methods of deriving such estimates. These methods, representing the simplest solutions to the problem, are studied below.

1.5.1 Ordinary Least Squares

Ignoring the fact that the M.A. parameters might be required for optimal prediction, estimates of the parameters $(\beta_1,\ldots,\beta_p)$, $(\gamma_o,\ldots,\gamma_r)$ of equation (1.2) may be obtained by O.L.S. The properties of such estimates differ according to whether B(L) = 0 or not, so that we will consider both cases in turn.

(a) B(L) = 0

When there are no A.R. parameters it is known that the O.L.S. estimator of (1.2) is consistent but not efficient (e.g. see Malinvaud [71]) but what is of importance is the extent of the loss of efficiency sustained by the use of the estimator. A numerical evaluation of this for particular models will be the concern of this section.[15]

Recently R. Hall has applied Fourier techniques to the estimation of the linear regression model with autocorrelated disturbances

i.e. $y(t) = C(L)\, x(t) + u(t)$ (1.24)

where $E(u(t)) = 0$ and the covariance matrix of u(t) is Ω. Taking the Fourier Transform of (1.24) and minimizing the sum of squares in the frequency domain with respect to the element of C(L) Hall finds that,

> "In a model with a stationary (time-invariant) disturbance process, the minimum variance unbiassed estimates of the distributed lag parameters (assuming they enter linearly) are given by weighted least squares in the frequency domain;

[15] At the same time this will enable an enumeration of the benefits of employing an efficient estimator.

the weights are the inverses of the spectrum of the disturbances" [38 p.31].

Utilizing this result (and assuming $C(L) = \gamma$ i.e. there are no lagged x(t) variables), Hall shows that the variances of the O.L.S. estimator $V(\hat{\gamma}_{OLS})$ and the efficiently weighted least squares (E.W.L.S.) estimator $V(\hat{\gamma}_{EWLS})$ are

$$V(\hat{\gamma}_{OLS}) = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{xx}(\lambda)\, f_{uu}(\lambda)d\lambda}{\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{xx}(\lambda)d\lambda\right]^2} \qquad (1.25)$$

$$V(\hat{\gamma}_{EWLS}) = \frac{1}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{f_{xx}(\lambda)}{f_{uu}(\lambda)}\, d\lambda} \qquad (1.26),$$

where $f_{xx}(\lambda)$, $f_{uu}(\lambda)$ represent the ordinates of the spectral density functions associated with x(t) and u(t) respectively, and the integration is performed over the range $-\pi \leq \lambda \leq \pi$.

The loss in efficiency is given by the ratio of $V(\hat{\gamma}_{EWLS})$ to $V(\hat{\gamma}_{OLS})$ and this ratio (Φ) will be called the <u>efficiency index</u>.

i.e. $\Phi = \dfrac{V(\hat{\gamma}_{EWLS})}{V(\hat{\gamma}_{OLS})}$

Although the efficiency index is a function of the spectral shape attributable to the x(t) and u(t) variables a lower bound to Φ may be established by an application of the following theorem:[16]

---

16 A proof of which may be found in Hannan [47 p.111].

Let $\phi_1 \ldots \phi_N$ be the N eigenvalues of Ω. Provided that N is large, $f_{uu}(\lambda)$ is not too irregular and the eigenvalues may be partitioned into two distinct sets with the property that $\phi_1 \simeq \phi_2 \simeq \ldots \simeq \phi_k$; $\phi_{N-k+1} \simeq \phi_{N-k+2} \simeq \ldots \simeq \phi_N$ where k is small relative to N, the efficiency index is given by

$$\Phi \geq \frac{4}{2 + \frac{\phi_1}{\phi_N} + \frac{\phi_N}{\phi_1}} \qquad (1.27).$$

As the quotation from Hall (cited above) stresses the covariance matrix of the Fourier-transformed disturbances is asymptotically diagonal with the spectral density ordinates as the non-zero elements. It may also be shown that the eigenvalues of the covariance matrix $\phi_1 \ldots \phi_N$ lie on this diagonal so that there is an equality between the two sets allowing the replacement of ϕ_1 and ϕ_N by $f_{uu}(\lambda)_{MAX}$ and $f_{uu}(\lambda)_{MIN}$ - respectively the maximum and minimum values of the spectrum.

Hall goes on to evaluate some efficiency losses (and the lower bound) under the assumption that x(t) and u(t) are generated by

$$x(t)\ (1-\delta L) = \varepsilon(t)$$
$$u(t)\ (1-\rho L) = e(t) \qquad (1.28).$$

Table 1.1 reproduces Hall's results for various values of δ and ρ when the number of observations is taken to be 75.

Table 1.1

Efficiency Index for Varying Values of the Parameters of Equation (1.24)

	ρ=0	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9
Lower bound	1	0.698	0.361	0.116	0.011
δ=0	1	0.835	0.600	0.342	0.105
δ=0.5	1	0.851	0.600	0.311	0.079
δ=0.8	1	0.914	0.714	0.389	0.083
δ=0.9	1	0.951	0.813	0.503	0.104
δ=0.95	1	0.974	0.894	0.649	0.149
δ=0.99	1	0.998	0.991	0.956	0.625

Source : Hall [38 p.37].

A notable feature of Table 1.1 is the severe overestimate of the loss in efficiency from O.L.S. provided by the least bound implying that use of this formula may be quite misleading if some knowledge of the spectral shapes of x(t) and u(t) is available. As there is now a considerable body of literature concerned with the spectral shapes of economic variables it is important to use this information in gaining an appreciation of the likely efficiency losses in the estimation of realistic economic models. Such criteria underlie the rationale for the filters summarized in (1.28) as both variables have Granger's "typical spectral shape" [33] which is commonly found by investigators working with yearly time series.

Table 1.1 leads Hall to the conclusion"....that the practical importance of autocorrelation in distributed lag models has been somewhat exaggerated. In the leading case $\rho = 0.5$ and $\delta = 0.95$, the

potential gain in efficiency from an autoregressive transformation of the variables....is about 10%" [38 p.38]. He recognizes that this is a product of the spectral shapes employed - particularly the large concentration of power at the origin for the x(t) variable in contrast to the u(t) variable - but is correct in arguing that such spectra are typical of equations estimated from yearly data.

Three objections may be levelled against Hall's methodology.

(i) Although the expression for Φ is ostensibly for the general distributed lag, in Hall's examples no such lag distribution exists. Denoting $z(t) = C(L)\, x(t)$ the variances should be constructed from the spectrum of z(t) (i.e. $f_{zz}(\lambda_j)$ replaces $f_{xx}(\lambda_j)$ in (1.25) and (1.26)) and as this is a product of the squared gain of the response function $C(\lambda_j)$ and the spectrum of x(t) it is tenable that even though $f_{xx}(\lambda_j)$ has the "typical spectral shape", $f_{zz}(\lambda_j)$ may not.[17]

(ii) As mentioned earlier the desire (and necessity) to incorporate distributed lag phenomena into econometric models intensified with the extensive use of quarterly data. Even though it is possible to reduce the spectra of x(t) to the first order shape given above by the use of deseasonalized series, it remains true that if u(t) incorporates omitted factors there is logic in requiring $f_{uu}(\lambda_j)$ to have peaks corresponding to the harmonics of a seasonal pattern.

---

[17] See Fishman [26] for a definition of the concepts of "frequency response function" and "squared gain". The convention is adopted that the response function associated with C(L) will be written as $C(\lambda_j)$ and the squared gain $|C(\lambda_j)|^2$. Therefore $|C(0)|^2$ is the gain at $\lambda_j = 0$.

Confirmation of this may be found in the article by Thomas and Wallis which demonstrates the existence of serial correlation in the residuals (at the seasonal frequencies), even though the data had been adjusted for a constant seasonal pattern (dummy variables) and a first order polynomial trend.

(iii) The efficiency index measures the <u>efficiency</u> of the O.L.S. estimator, but not the <u>efficacy</u> of the O.L.S. <u>formula</u> when serial correlation is present. It is easily shown that the variance of $\hat{\gamma}_{OLS}$ given by the O.L.S. formula is

$$V^*(\hat{\gamma}_{OLS}) = \frac{1}{\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{xx}(\lambda)\, i\lambda\right]} \qquad (1.29).$$

A distinction between (1.25) and (1.29) is essential for any judgement concerning the robustness of O.L.S. Most computer programs calculate the variance of $\hat{\gamma}_{OLS}$ from (1.29) which is not the <u>true</u> variance. This enables us to see the dilemma facing the O.L.S. proponent. On the one hand it appears that O.L.S. is not greatly inefficient but the well known O.L.S. formula for the variance is not exact, and on the other in order to calculate the exact variance knowledge of $f_{uu}(\lambda_j)$ is required. If such knowledge exists the E.W.L.S. estimator might just as well have been used.

The three objections enumerated above suggest a further inquiry into the performance of O.L.S. but with the following items given particular attention.

(i) The extent to which the O.L.S. formula is incorrect.

(ii) The efficiency losses under a seasonal specification of the disturbances and/or exogenous variables.

(iii) The impact of a more realistic distributed lag upon the efficiency of the estimator.

Table 1.2 presents the transfer functions that were used for research into the three factors described above.

Table 1.2

Alternative Schemes to Generate x(t), C(L) and u(t)

C(L)	x(t)	u(t)
$C(L) = c$	$x(t)(1-0.9L)=\varepsilon(t)$	$u(t)=(1+\alpha_1 L)e(t)$
$C(L) = c(1+2L+3L^2+4L^3+5L^4+6L^5+6L^6+5L^7+4L^8+3L^9+2L^{10}+L^{11})$	$x(t)(1-0.9L)(1-0.8L^4)=\varepsilon(t)$	$u(t)=(1+\alpha_1 L)(1+\alpha_4 L^4)e(t)$

A short commentary on the parameters given in Table 1.2 is in order. Firstly, the three sets of values for α_1 and α_4 were taken as

$\alpha_1 = 0.4, 0.6, 0.8$

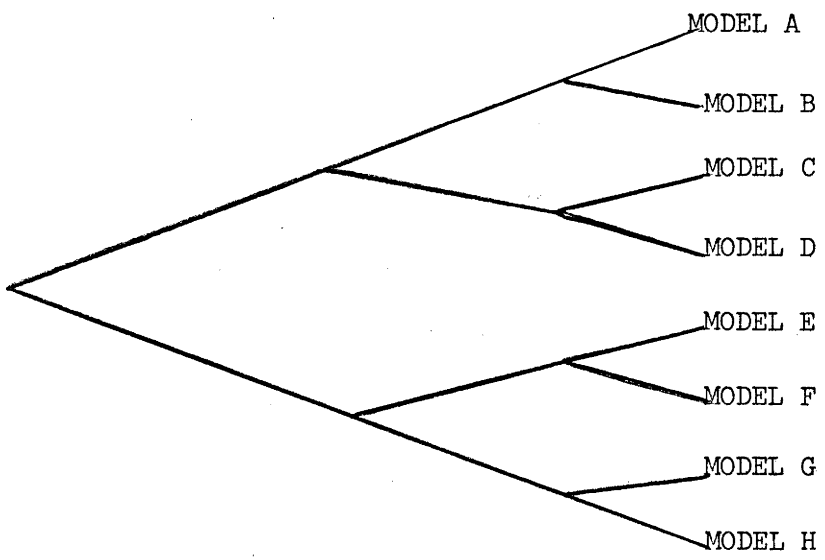
$(\alpha_1, \alpha_4) = (0.4, 0.2), (0.6, 0.3), (0.8, 0.3)$

with the expectation that these will reproduce the effects associated with low, medium, and high autocorrelation in the disturbances. Secondly, the seasonal pattern given for x(t) was adapted from Nerlove and Grether [36]. Lastly, the second form of C(L) is the "inverted V" lag distribution first proposed by F. de Leeuw [67]. This lag distribution, with its peak at six time periods and zero impact after twelve, has a shape characteristic of that found in a number of empirical studies, and provided a good fit to investment expenditure data.

Schematically Table 1.2 may be viewed as the decision tree of Figure 1.1. Each node corresponds to a new column so that there are three nodes in the tree. As each column is composed of two alternative specifications there will be eight distinct models and these are labelled A - H.

Figure 1.1

Schematic Representation of the Models Implicit in Table 1.2



As a guide to the interpretation of Figure 1.1 we record that Model B has the following generating functions : $C(L) = c$; $x(t)(1-0.9L) = \varepsilon(t)$; $u(t) = (1+\alpha_1 L)(1+\alpha_4 L^4)e(t)$.

Table 1.3 sets out the efficiency indexes Φ and ψ (the latter being the ratio of (1.26) to (1.29)) for Models A-H classified by the extent of autocorrelation in the residuals.

Table 1.3

Efficiency Indexes of O.L.S. for Models A-H

Model	Φ			ψ		
	Low	Med.	High	Low	Med.	High
A	0.949	0.878	0.718	1.785	2.143	2.211
B	0.880	0.745	0.615	2.174	2.734	2.850
C	0.982	0.944	0.790	1.930	2.387	2.524
D	0.977	0.933	0.780	2.694	3.928	4.152
E	0.999	0.999	0.999	1.953	2.549	3.225
F	0.993	0.988	0.988	2.695	4.065	5.143
G	0.999	0.999	0.999	1.958	2.558	3.237
H	0.999	0.999	0.999	2.795	4.273	5.408

Some useful norms may be extracted from Table 1.3.

1. The loss in efficiency of O.L.S. (as measured by Φ) is only once greater than 15% in the medium autocorrelation range but can become substantial if there is a good deal of autocorrelation. Overall however the results support Hall's sanguine attitude towards the robustness of O.L.S.

2. Dichotomizing Table 1.3 according to the complexity of the lag distribution (i.e. Models (A-D) and (E-H)) it is striking that O.L.S. is far more efficient with the de Leeuw lag form. An explanation lies in the magnitude of the squared gain of C(L). There are two major peaks (at roughly $2\pi/5$ and $3\pi/5$) with $|C(0)|^2$, $|C(\frac{2\pi}{5})|^2$, $|C(\frac{3\pi}{5})|^2$ being in the (approximate) ratios 2800 : 7 : 1 so that the product of the spectrum of x(t) and the squared gain function $|C(\lambda)|^2$ will cause $f_{zz}(\lambda)$ to have a concentration of power at the origin - precisely the same

circumstance in which Hall found that O.L.S. was efficient. From this example we conclude that intricate lag patterns may actually increase the efficiency of O.L.S.[18]

3. The effects of peaks in $f_{uu}(\lambda)$ when $f_{xx}(\lambda)$ has none may be evaluated by a comparison of Models (B,D) and (F,H). As expected efficiency losses are greatest when the spectra are dissimilar but in contrast the formula variance understates the true variance by a larger amount when peaks are present in both $f_{uu}(\lambda)$ and $f_{xx}(\lambda)$ (compare especially Models B and D), and this tendency becomes more emphatic as the degree of autocorrelation in the disturbances rises.

4. What might be termed the efficacy index (ψ) illustrates quite dramatically the extent to which the traditional O.L.S. variance formula can be misleading and makes one aware of the failure of this statistic to provide a reliable guide to the acceptance and rejection of hypotheses. No solace is to be gained by O.L.S. users from Table 1.3 as even in those cases where O.L.S. is efficient the formula variance is at best five eights of the true value.

(b) $C(L) = 0$

Having established some idea of the loss of efficiency in employing O.L.S. when there are lagged exogenous variables among the regressors and autocorrelation in the disturbances, there is another context in which autocorrelation induces non-optimal properties into this estimator

---

[18] Only one-parameter models were examined in order to avoid the difficulties associated with a suitable definition of efficiency in multi-parameter cases.

i.e. when a lagged dependent variable is among the regressors. Under such conditions it is known that the O.L.S. estimator is not consistent. In what is to follow closed form expressions are given for the inconsistency of O.L.S. and the asymptotic variance of E.W.L.S. under the assumption that the autoregression is of first order i.e. of the Koyck type.[19]

The ARMA(1,q) model was selected.

i.e. $(1-\beta L)\, y(t) = A(L)\, \varepsilon(t)$ (1.30).

Letting $B(\lambda)$, $A(\lambda)$ stand for the frequency response functions associated with $(1-\beta L)$ and $A(L)$, Appendix 2 shows that the inconsistency of $\hat{\beta}_{OLS}$(IC), the variance of $\hat{\beta}_{EWLS}$($V(\hat{\beta}_{EWLS})$), and the formula variance for $\hat{\beta}_{OLS}$ ($V^*(\hat{\beta}_{OLS})$) are

$$IC(\hat{\beta}_{OLS}) = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|A(\lambda)|^2}{|B(\lambda)|^2}\, d\lambda \right]^{-1} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|A(\lambda)|^2 e^{i\lambda}}{B(\lambda)}\, d\lambda \right) \tag{1.31}$$

$$V(\hat{\beta}_{EWLS}) = \frac{1}{\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|B(\lambda)|^2}\, d\lambda \right]} \tag{1.32}$$

$$V^*(\hat{\beta}_{OLS}) = \frac{1}{\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} f_{yy}(\lambda) d\lambda \right]} \tag{1.33}.$$

The expressions of equations (1.31) - (1.33) were evaluated for a number of different values of β and the two specifications of $A(L)$

---

19 Unfortunately it is not possible to derive the asymptotic variance of the O.L.S. estimator for this case but in practice the necessity of knowing $f_{uu}(\lambda)$ makes this rather academic. As it is the formula variance that will be quoted in most studies we wish to evaluate the extent to which this may be misleading as an indicator of the E.W.L.S. solution.

used previously – the parameter combinations to represent degrees of autocorrelation being α_1 = (0.4, 0.6, 0.8), (α_1, α_4) = (0.4,0.2), (0.6, 0.2),(0.8, 0.2). Tables 1.4 and 1.5 set out the efficacy index ψ and the absolute value of the inconsistency classified by the extent of autocorrelation in the residuals.

Table 1.4

Efficacy Index for and Inconsistency of O.L.S.

$A(L) = (1+\alpha_1 L)$.*

β	ψ			Inconsistency		
	Low	Med.	High	Low	Med.	High
0.1	1.24	1.48	1.80	0.319	0.401	0.440
0.2	1.32	1.60	1.96	0.291	0.360	0.392
0.3	1.40	1.72	2.12	0.260	0.317	0.343
0.4	1.48	1.84	2.28	0.227	0.274	0.295
0.5	1.56	1.96	2.44	0.192	0.230	0.246
0.6	1.64	2.08	2.60	0.156	0.185	0.197
0.7	1.72	2.20	2.76	0.119	0.139	0.148
0.8	1.80	2.32	2.92	0.080	0.093	0.099
0.9	1.88	2.56	3.08	0.040	0.047	0.049

* The integrals were approximated by summation over 100 points.

Table 1.5

Efficacy Index for and Inconsistency of O.L.S.

$$A(L) = (1+\alpha_1 L)(1+\alpha_4 L^4)$$

β	ψ			Inconsistency		
	Low	Med.	High	Low	Med.	High
0.1	1.29	1.54	1.87	0.320	0.402	0.441
0.2	1.37	1.67	2.04	0.294	0.364	0.395
0.3	1.46	1.80	2.20	0.267	0.325	0.351
0.4	1.56	1.95	2.41	0.240	0.287	0.307
0.5	1.68	2.11	2.63	0.210	0.247	0.263
0.6	1.81	2.30	2.88	0.178	0.206	0.218
0.7	1.98	2.54	3.19	0.143	0.162	0.170
0.8	2.20	2.84	3.57	0.101	0.113	0.118
0.9	2.47	3.21	4.06	0.054	0.059	0.062

Once again we notice that even if O.L.S. were fully efficient the formula variance would be misleading. However what is more disturbing (and emphasizes that preoccupation with efficiency may distract attention from other properties) is the size of the inconsistency displayed. Even for small amounts of autocorrelation this may be quite substantial and mitigates against the use of O.L.S. if a consistent estimator is available. At first glance the values for inconsistency follow what seems to be an unusual pattern i.e. as the autoregressive parameter rises the inconsistency falls. The explanation resides in the fact that the greater the concentration of spectral mass at the origin the closer will O.L.S. be to the M.L. estimate (see Section 4.4 for more on this), but another justification is obtained from considering the inconsistency of β_1 when the disturbances follow a first order

autoregression with parameter ρ viz. $\text{plim}(\hat{\beta}_1-\beta_1) = \frac{\rho(1-\beta_1^2)}{1+\beta_1\rho}$

(Malinvaud [71 p.460]). By differentiating this with respect to β_1 it is easily seen that the conclusion holds in this model as well. Finally the tables show that the peak in the disturbance spectrum at $\pi/2$ causes higher inconsistency etc., and because the peak is not very large suggests that O.L.S. should be regarded with greatest caution when monthly and quarterly data are being modelled, as the possibility of peaked disturbance spectra is much greater in such cases.

1.5.2. Generalized Least Squares(G.L.S.)

Relying on the well established theorem that the G.L.S. estimator of the parameters of the general linear model is maximum likelihood (M.L.) under the restriction that the disturbances are n.i.d. with covariance matrix Ω, one is tempted to argue for the use of G.L.S. in place of O.L.S. when faced with the need to estimate an ARMA/ARMAX model.

A theoretical objection to G.L.S. is that knowledge of the true covariance matrix (which is required) is rarely available. An early suggestion was to replace Ω by a consistent estimator $\hat{\Omega}$ but it is now known that the estimator obtained with $\hat{\Omega}$ is asymptotically equivalent to that resulting from the use of Ω iff plim $Z^Tu = 0$ where Z is a (N x K) matrix of K regressors and u is the (N x 1) vector of disturbances ($u(t) = A(L)e(t)$).[20] If the regressors are truly exogenous then this condition holds and G.L.S. is appropriate. Unfortunately ARMA/ARMAX models are characterized by the presence of lagged dependent variables in the Z matrix and the necessary condition will not be met.

[20] Zellner [123] appears to have been the first to state the general theorem.

There are further objections to this form of G.L.S. i.e. based on $\hat{\Omega}$. One is that incorporation of information concerning the order of the M.A. into the estimator will generate estimates that are asymptotically more efficient than G.L.S. If no preconceptions concerning the error process exist G.L.S. will be optimal, but the computational difficulties raised by the necessity to invert an (N x N) matrix are formidable and - more to the point - with a finite sample the autocorrelations computed from the residuals for use in $\hat{\Omega}$ would be very unreliable at high lags of the correlogram.

1.5.3. Correlogram Estimates

For completeness some consideration must be given to estimates of M.A. parameters derivable from the autocovariance function of the residuals (which in practice would be replaced by consistent estimates). For an ARMA (p,1) model this would be accomplished in two steps:-

(a) Estimate $\beta_1,\ldots,\beta_p$ by some consistent estimator e.g. Liviatan's [69] with $y(t-q-1),\ldots,y(t-q-p)$ as instruments for $y(t-1),\ldots,y(t-p)$. The resulting estimates $\hat{\beta}_1,\ldots,\hat{\beta}_p$ are consistent.

(b) The first two covariances of the M.A. are

$$\gamma(0) = \sigma^2(1+\alpha_1^2)$$

$$\gamma(1) = \sigma^2\alpha_1$$

so that by forming estimates $\hat{\gamma}(0)$, $\hat{\gamma}(1)$ (from $\hat{\beta}_1,\ldots,\hat{\beta}_p$) we may obtain an estimate $\hat{\alpha}_1$ from these two equations. Walker [110] has shown that these estimates are consistent but Whittle [115] calculated that if $\alpha_1 = 0.5$ the asymptotic variance was 3.8 times that of the M.L. estimate. Efficiency losses of this magnitude

make such estimates of little value for hypothesis testing but they will be used in later analysis when only a consistent estimator is required.

1.6 Conclusion

Three topics have been highlighted in this chapter.

(i) ARMA/ARMAX models are likely to arise when modelling single time series or explaining dynamic relations between multiple time series.

(ii) It is worthwhile to obtain estimates of the parameters of these models for the dual purpose of validating hypotheses and best prediction.

(iii) Traditional estimation procedures such as O.L.S. are unlikely to realize the potential gains from (ii) in the sample sizes commonly available with quarterly data.

If one accepts the above contentions it is natural to develop estimators to realize any large sample benefits, and then to determine the comparative rankings of all estimators in small samples. It is to the first of these tasks that we turn in Chapter 2 and the second in Chapters 3 and 4.

Appendix 1

Time Aggregation and ARMA Models

Only the case when $y(t)$ is a stock variable will be considered but this is illustrative of the general approach. Let $B(L)y(t) = e(t)$ be a p'th order A.R. indexed on $t = 1,\ldots,N$. Assume that a new series is formed from every k'th observation i.e. $r = k, 2k,\ldots$ and that in the translation from the t domain the order of the A.R. is preserved i.e. $B(L)y(r) = \varepsilon(r)$. This means that written in the t domain this will be $W(L)y(t) = e(t)$ where $W(L)$ will have non-zero coefficients at $L^{o}, L^{k}, L^{2k},\ldots$ only e.g. $W(L)$ will be (1.12) of the text. It is easily seen that $W(L)$ must be of order pk and it is necessary to seek the polynomial $T(L)$ that transforms $B(L)$ into $W(L)$.

Lemma: The polynomial $T(L)$ in the product $T(L)B(L) = W(L)$ has order $p(k-1)$.

Proof: Since $W(L)$ is of order pk and $B(L)$ of order p, $T(L)$ must be of order $p(k-1)$.

Theorem: The polynomial $T(L)$ in the product $T(L)B(L) = W(L)$ is unique.

Proof: There are $\ell = p(k-1)+1$ elements in $T(L)$ i.e. including the parameter attached to L^{o} and the elements of all polynomials may be arranged in matrix form as

$$\begin{bmatrix} 1 & & & \\ \beta_1 & 1 & & \\ & \beta_1 & .1 & \\ \beta_p & & 1 & \\ 0 & \beta_p & & \beta_1 \quad 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ t_1 \\ \cdot \\ \cdot \\ t_\ell \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 1 \\ w_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ w_{pk} \end{bmatrix}$$

i.e. $B\,t = w$

so that a non-trivial and unique solution for t will exist if B is non singular. The form of B guarantees that det (B) = 1 so that a solution for t exists and it is unique.

Now consider the original equation

$$B(L)y(t) = e(t)$$

$$\therefore \quad T(L)B(L)y(t) = T(L)e(t) = u(t).$$

Assume that there is q'th order autocorrelation in the r domain. Then this implies q k order autocorrelation in the t-domain. As the maximum order of T(L) in the t domain is p(k-1) the order of the auto-correlation in the r domain can be derived from

$$q\ k = p(k-1),$$

and is

$$q = \frac{p(k-1)}{k} \quad .$$

A difficulty with this expression arises in that it may be fractional whereas q must be an integer, but this is resolved by noting that fractional lags will always be truncated to the integer immediately below their exact value when only discrete data is available e.g. a $3\frac{1}{2}$ period lag must become a 3 period lag in discrete time. Therefore time transforms will change an ARMA (p,o) model to ARMA $(p,I(\frac{p(k-1)}{k}))$ which is the formula contained in the text.

Appendix 2

Derivation of Equations (1.31)-(1.33) of the Text

Consider the first order A.R.

$$y(t) = \beta\, y(t-1) + u(t) \qquad \text{(A2.1)}.$$

The O.L.S. estimate is

$$\hat{\beta}_{OLS} = \frac{\sum_{t=1}^{N} y(t)y(t-1)}{\sum_{t=1}^{N} y^2(t-1)}$$

$$= \frac{\sum_{t=1}^{N} y(t-1)[\beta y(t-1)+u(t)]}{\sum_{t=1}^{N} y^2(t-1)}$$

$$= \beta + \frac{\frac{1}{N}\sum_{t=1}^{N} y(t-1)u(t)}{\frac{1}{N}\sum_{t=1}^{N} y^2(t-1)}$$

Now $\underset{N\to\infty}{\text{plim}}\ \hat{\beta}_{OLS} = \beta + \dfrac{\underset{N\to\infty}{\text{plim}}\ \frac{1}{N}\sum_{t=1}^{N} y(t-1)u(t)}{\underset{N\to\infty}{\text{plim}}\ \frac{1}{N}\sum_{t=1}^{N} y^2(t-1)}$

so that the extra term measures the inconsistency of O.L.S. It is well known that

$$\underset{N\to\infty}{\text{plim}}\ \frac{1}{N}\sum_{t=1}^{N} y(t-1)u(t) = \gamma_{yu}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yu}(\lambda)\, e^{i\lambda}\, d\lambda$$

$$\underset{N\to\infty}{\text{plim}}\ \frac{1}{N}\sum_{t=1}^{N} y^2(t-1) = \gamma_{yy}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yy}(\lambda)\, d\lambda$$

∴ The inconsistency may be written as

$$\underset{N\to\infty}{\text{plim}} \ (\hat{\beta}_{OLS} - \beta) = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yy}(\lambda)d\lambda\right]^{-1}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yu}(\lambda)e^{i\lambda}d\lambda\right] \qquad (A2.2).$$

Now equation (A2.1) may be written in the frequency domain as

$$f_{yu}(\lambda)B(\lambda) = f_{uu}(\lambda)$$

where $B(\lambda) = 1 - \beta e^{i\lambda}$.

Two cases are distinguished

(a) $\underline{u(t) = \varepsilon(t)}$

The error term is white noise and $f_{uu}(\lambda) = \frac{\sigma^2}{2\pi}$ so that (A2.2) becomes

$$\text{I.C.} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yy}(\lambda)d\lambda\right]^{-1}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma^2 e^{i\lambda}}{2\pi B(\lambda)}\,d\lambda\right] \qquad (A2.3).$$

From the theory of residues $\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma^2 e^{i\lambda}}{2\pi B(\lambda)}\,d\lambda = 0$ so that

I.C. = 0 as is expected.

(b) $\underline{u(t) = A(L)\varepsilon(t)}$

Equation (A2.2) will become

$$\text{I.C.} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yy}(\lambda)d\lambda\right]^{-1}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma^2|A(\lambda)|^2 e^{i\lambda}}{2\pi\, B(\lambda)}\,d\lambda\right]$$

or using $f_{yy}(\lambda) = \frac{\sigma^2}{2\pi}\,\frac{|A(\lambda)|^2}{|B(\lambda)|^2}$

$$\text{I.C.} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|A(\lambda)|^2}{|B(\lambda)|^2}\,d\lambda\right]^{-1}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma^2|A(\lambda)|^2 e^{i\lambda}}{2\pi\ B(\lambda)}\,d\lambda\right] \qquad (A2.4).$$

The formula variance for O.L.S. will obviously be

$$V^*(\hat{\beta}_{OLS}) = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} f_{yy}(\lambda)d\lambda \right]^{-1} \qquad \text{(A2.5)},$$

and for E.W.L.S.

$$V(\hat{\beta}_{EWLS}) = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{f_{yy}(\lambda)}{f_{uu}(\lambda)} d\lambda \right]^{-1}$$

$$= \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|B(\lambda)|^2} d\lambda \right]^{-1} \qquad \text{(A2.6)}.$$

Equations (2.4), (2.5) and (2.6) are (1.31) - (1.33) of the text.

CHAPTER 2. The Estimation of ARMAX Models

2.1 Introduction

Chapter 1 has established a potential need for econometricians to consider ARMAX models when presented with the task of inferring economic relationships from time series data and some mention has been made of possible estimators of the parameters of these equations. For a variety of reasons the estimators mentioned were rejected as unsuitable thereby emphasizing the importance of designing new estimators which have the desirable properties (mentioned in Chapter 1) of consistency and efficiency. Because of the association of such characteristics with maximum likelihood estimators it is natural to propose the maximization of the likelihood stemming from ARMAX equations.

As a prelude to the derivation of these estimators it is requisite to summarize some concepts from the numerical methods literature and Section 2.2 is devoted to such a resumé. Following this three estimators that maximize the likelihood - two in the time domain and one in the frequency domain - are outlined and a framework is developed that will enable an evaluation (in Section 2.5) of other estimators suggested in the literature. A distinction is drawn throughout between the time and frequency domain aspects of the problem.

In attempting to translate the theoretical constructs of Sections 2.3 and 2.4 into computable algorithms a number of difficulties arose and these are outlined (along with the solutions adopted) in Section 2.6. Lastly a survey of the existing statistical theory that may be invoked to justify usage of these estimators is given, and this identifies those areas in which such theory is deficient. These deficiencies become the subject of the research presented in Chapter 3.

2.2 Non-Linear Algorithms

As the estimators to be proposed will involve functional minimization, reference must be made to various techniques available for optimizing functions. There are a proliferation of these (see Powell [92] for a

recent survey) but this thesis will be concerned exclusively with the class known as "gradient" methods - for the dual reasons of simplicity and a long and successful history of use.[1] Three members of this family assume particular importance for the thesis.

(i) Steepest-Descent

(ii) Gauss-Newton

(iii) Newton-Raphson

As a good discussion of the derivation of each is now readily available to econometricians in Malinvaud [71] only the general algorithmic form will be mentioned.

Denote the function to be optimized by S which may be regarded as a scalar formed from the premultiplication of an (N×1) column vector by its transpose (i.e. $S = e^T e$ where T denotes transposition), and assume that e is a function of a (K×1) parameter vector θ. Then the following iterative formulae are prescribed for (i), (ii) and (iii) respectively:

$$\text{(i)} \quad \theta^{(n)} - \theta^{(n-1)} = -\frac{\partial S}{\partial \theta}$$

$$\text{(ii)} \quad \theta^{(n)} - \theta^{(n-1)} = -\left(\frac{\partial e^T}{\partial \theta} \frac{\partial e}{\partial \theta} \right)^{-1} \frac{\partial e^T}{\partial \theta} e \qquad (2.1)$$

$$\text{(iii)} \quad \theta^{(n)} - \theta^{(n-1)} = -\left(\frac{\partial^2 S}{\partial \theta \partial \theta^T} \right)^{-1} \frac{\partial S}{\partial \theta}$$

where $\theta^{(n)}$ represents the estimate of the parameter vector at the n'th iteration and all derivatives are evaluated with estimates from the preceeding ((n-1)'th)iteration.

A number of points must be emphasized in connection with (2.1).

[1] Actually some experimentation with alternative algorithms was conducted at an early stage in the research but this was discontinued when there was substantial evidence that gradient methods would be satisfactory.

(a) The Gauss-Newton formula is only applicable if the row-column vector factorization exists but neither (i) nor (iii) are restricted to this functional form.

(b) If e is a linear function of θ then S is a quadratic in θ and convergence will be achieved in one step from any initial value $\theta^{(0)}$. Furthermore if S is the sum of squares function corresponding to a linear regression, setting $\theta^{(0)} = 0$ reduces (ii) to the O.L.S. formula.

(c) If S has the row-column vector factorization then $2 \frac{\partial e^T}{\partial \theta} e = \frac{\partial S}{\partial \theta}$ so that (i) - (iii) become members of a family of algorithms with the general structure:

$$\theta^{(n)} - \theta^{(n-1)} = -B^{-1} \frac{\partial S}{\partial \theta} \qquad (2.2)$$

where we may by making appropriate substitutions for B (referred to as the weighting matrix) generate (i) - (iii) e.g. B = I yields Steepest Descent.

2.3 Time Domain Estimation of ARMAX Models

The ARMAX equation to be estimated is given by (1.2) and is reproduced below as (2.3)

$$B(L)\, y(t) = C(L)\, x(t) + u(t) \qquad (2.3).$$

$$u(t) = A(L)\, e(t)$$

From Chapter 1 the model to be estimated has lags of (p,r,q) making it necessary to delete m observations (m = maximum of p and r) from the series. These omissions are termed pre-period values and will be distinguished by negative indexes e.g. y(-1), y(-2) etc. Under this

convention, if there were initially G observations, N = G-m will remain and these are arranged in (1×N) vectors {y(0)...y(N-1)}, {x(0)...x(N-1)} etc. Such an enumeration, albeit somewhat unusual, is convenient in that it reserves negative indexes for pre-period values.

Following Phillips [90], the model may be arranged in matrix notation as

$$y = Y\beta + X\gamma + u \tag{2.4}$$

$$u = Q\,\bar{e} \tag{2.5}$$

where

$$y = \begin{bmatrix} y(0) \\ \vdots \\ y(N-1) \end{bmatrix} \quad Y = \begin{bmatrix} y(-1) & \cdots & y(-p) \\ \vdots & & \vdots \\ y(N-2) & \cdots & y(N-p-1) \end{bmatrix} \quad X = \begin{bmatrix} x(0) & \cdots & x(-r) \\ \vdots & & \vdots \\ x(N-1) & \cdots & x(N-r-1) \end{bmatrix}$$

$$u = \begin{bmatrix} u(0) \\ \vdots \\ u(N-1) \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \gamma = \begin{bmatrix} \gamma_o \\ \vdots \\ \gamma_r \end{bmatrix} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{bmatrix}$$

$$Q = \begin{bmatrix} \alpha_q & \cdots & 1 & & \\ & \ddots & & \ddots & \\ & & \ddots & & 0 \\ 0 & & \ddots & & \\ & & \alpha_q & \cdots & 1 \end{bmatrix} \qquad \bar{e} = \begin{bmatrix} e(-q) \\ \vdots \\ e(0) \\ \vdots \\ e(N-1) \end{bmatrix}$$

The matrices and vectors are of the following orders:

y - Nx1	γ - (r+1)x1	X - Nx(r+1)
Y - Nxp	β - px1	Q - Nx(N+q)
$\bar{e}$ - (N+q)x1	u - Nx1	α - qx1,

so that all products are conformable.

Partitioning (2.5) as indicated by the broken lines u may be rewritten as

$$u = Me + M^*e^* \qquad (2.6).$$

A direct comparison of (2.5) and (2.6) reveals that

$$Q = [M^* : M]$$

$$\bar{e} = \begin{bmatrix} e^* \\ \cdots \\ e \end{bmatrix}$$

and M^* is (Nxq), e^* is (qxl), M is (NxN), and e is (Nxl).

Under the assumption that all pre-period values are constants and that e(t) is n.i.d. Phillips derives the likelihood of {y(0).....y(N-1)} from the joint density function of {e(0)....e(N-1)} as

$$L = F(y/x,y^*,x^*;\ \theta,\sigma^2) = (2\pi)^{-N/2}(\sigma^{-2})^{N/2}\ \exp\{-\tfrac{1}{2}\sigma^{-2}e^Te\} \qquad (2.7).$$

Here y*,x* and e* (when transposed) are the vectors {y(-1)....y(-p)}, {x(-1)....x(-r)}, {e(-1)....e(-q)} and the notation signifies that the likelihood is conditional upon these as constants.

Alternatively the likelihood of {y(0)....y(N-1)} might be derived from the joint density of {u(0)....u(N-1)}. To do this we recognize that the covariance matrix of u(obtained from (2.5)) is

$$E(uu^T) = \Omega = E(Q\ \bar{e}\ \bar{e}^T\ Q^T)$$

$$\Omega = \sigma^2 Q\ Q^T \qquad (2.8),$$

and using this and the knowledge that u(t) is a linear combination of n.i.d. variables e(t)....e(t-q) the likelihood may be written

$$L = (2\pi)^{-N/2} \det \Omega^{-\frac{1}{2}} \exp \{-\frac{1}{2} u^T \Omega^{-1} u\} \qquad (2.9).$$

Equation (2.9) is of special interest in that maximization of the exponent will yield generalized least squares estimates (see Section (1.5)).

2.3.1 The Phillips Estimator

Employing the logarithmic form of L, discarding terms of lower order than the exponent, and forming S = -L, reduces (2.7) to

$$S = e^T e \qquad (2.10).$$

The necessary first order conditions for a minimum to (2.10) are

$$\frac{\partial S}{\partial \theta} = 0 \qquad (2.11).$$

$$\frac{\partial S}{\partial \sigma^2} = 0$$

In order to express e as a function of the parameters, (2.4) and (2.6) may be combined to yield

$$e = M^{-1} (y - Y\beta - X\gamma - M^* e^*) \qquad (2.12).$$

Identifying the parameters of the likelihood θ as $[\beta,\gamma,\alpha,e^*]$ and adopting the convention that

$$\delta = \begin{bmatrix} \beta \\ \cdots \\ \gamma \end{bmatrix} \qquad Z = \left[\begin{array}{c:c} Y & X \end{array}\right]$$

(2.12) may be rewritten as

$$e = M^{-1} (y - Z\delta - M^* e^*) \qquad (2.13).$$

The objective function (2.10) will be minimized by use of a Gauss-Newton algorithm which will require the following derivatives:

$$\begin{aligned}
\frac{\partial e}{\partial \delta_j} &= -M^{-1} Z_j && j = 1,\ldots,(p+r+1) \\
\frac{\partial e}{\partial \alpha_j} &= -M^{-1} \bar{E}_j && j = 1,\ldots,q \\
\frac{\partial e}{\partial e^*_j} &= -M^{-1} M^*_j && j = 1,\ldots,q
\end{aligned} \tag{2.14}$$

where the subscript j refers to the j'th column of a matrix e.g. Z_j would be the j'th column of Z and the (Nxq) matrix $\bar{E}$ has the form

$$\bar{E} = \begin{bmatrix} e(-1) \ldots\ldots e(-q) \\ \vdots \qquad\qquad \vdots \\ e(N-2) \ldots\ldots e(N-q-1) \end{bmatrix} .$$

The set of derivatives in (2.14) are used as inputs into the Gauss-Newton algorithm presented in Section 2.2 and this provides the sequence of iterations stated by Phillips in his paper.[2]

---

[2] Actually Phillips' version is slightly different because he estimates the linear combinations of $\{e(-1)\ldots e(-q)\}$ and $\{\alpha_1 \ldots \alpha_q\}$ formed from the matrix product M*e*. The present estimator may be more efficient. It is important to note that the estimating equations given by Phillips merely reflect the choice of the Gauss-Newton method to minimize (2.10) and a different choice would have lead to different equations. To some extent Phillips' derivation obscures this.

If the sequence of iterations converges then at the final iteration an estimate of the residual variance is given by

$$\hat{\sigma}^2 = e^T e/N \qquad (2.15a),$$

and an estimate of the covariance matrix of $\sqrt{N}(\hat{\theta}-\theta)$ by

$$V_{\hat{\theta}\hat{\theta}} = N\,\sigma^2 \left(\frac{\partial e^T}{\partial \theta} \quad \frac{\partial e}{\partial \theta} \right)^{-1} \qquad (2.15b).$$

2.3.2 An Aitken Estimator

Equation (2.9) gave an alternative likelihood for ARMAX processes that corresponds to the Aitken estimator when only the exponent is considered. Taking logs of this and omitting terms of lower order than the exponent the function to be minimized is[3]

$$S = u^T \Omega^{-1} u \qquad (2.16).$$

Owing to the difficulty of obtaining a vector factorization of S it is easiest to minimize S with a Newton-Raphson algorithm, the required first and second derivatives being

$$\frac{\partial S}{\partial \delta_j} = -2 Z_j^T \Omega^{-1} u \qquad j = 1,..,(r+p+1) \qquad (2.17)$$

$$\frac{\partial S}{\partial \alpha_j} = -u^T \Omega^{-1} \frac{\partial \Omega}{\partial \alpha_j} \Omega^{-1} u \qquad j = 1,..,q \qquad (2.18)$$

$$\frac{\partial^2 S}{\partial \delta_i \partial \delta_j} = 2 Z_j^T \Omega^{-1} Z_i \qquad i,j = 1,..,(r+p+1) \qquad (2.19)$$

$$\frac{\partial^2 S}{\partial \alpha_i \partial \delta_j} = 2 Z_j^T \Omega^{-1} \frac{\partial \Omega}{\partial \alpha_i} \Omega^{-1} u \qquad \begin{matrix} i = 1,..,q \\ j = 1,..,(r+p+1) \end{matrix} \qquad (2.20)$$

---

[3] Walker [114] has shown that $\lim_{N\to\infty} \det \Omega = 1$. It is feasible that one might express Ω as a function of α and σ^2 and include it in the function to be minimized, but the equations corresponding to (2.17)-(2.20) would be analytically intractable. Numerically it would be simpler but the size of sample considered in this thesis should be large enough for the limit to be operative.

$$\frac{\partial^2 S}{\partial\alpha_i \partial\alpha_j} = u^T \Omega^{-1} \frac{\partial\Omega}{\partial\alpha_i} \Omega^{-1} \frac{\partial\Omega}{\partial\alpha_j} \Omega^{-1} u + u^T \Omega^{-1} \frac{\partial\Omega}{\partial\alpha_j} \Omega^{-1} \frac{\partial\Omega}{\partial\alpha_i} \Omega^{-1} u$$

$$-u^T \Omega^{-1} \frac{\partial^2\Omega}{\partial\alpha_i \partial\alpha_j} \Omega^{-1} u \quad i, j = 1,..,q \qquad (2.21).$$

Equations (2.17) - (2.21) are used as inputs into the Newton-Raphson algorithm. If the sequence is convergent estimates of the residual variance and the covariance matrix of $\sqrt{N}(\hat{\theta}-\theta)$ is given at the final iteration by

$$\hat{\sigma}^2 = u^T \Omega^{-1} u/N \qquad (2.22)$$

$$V_{\hat{\theta}\hat{\theta}} = I^{-1}$$

$$\mathbf{I}(i,j) = \frac{2}{N} \frac{\partial^2 S}{\partial\theta_i \partial\theta_j} \qquad i, j = 1,..,(p+r+q+1) \qquad (2.23).$$

Any preference for this estimator - referred to as the Aitken Estimator in the Time Domain (A.T.D.) - is conditional on the reduction in the number of parameters to be estimated (owing to the deletion of e*) and the increased computational cost incurred by the inversion of Ω.

2.4 Frequency Domain Estimation of ARMAX Models

Recently there has been a revival of interest in the application of Fourier Analysis to econometric problems. This interest - centered on the spectral representation of a time series - has been stimulated by the demonstration that the extraction of information concerning (approximate) periodicities etc. from economic series is performed more easily with the spectral density function than its time domain counterpart, the autocorrelation function. As the two are Fourier Transform pairs they contain equivalent amounts of information but the ability to access

such information differs.

Owing to the simple form of the covariance matrix of autocorrelated disturbances in the frequency domain (and conversely its complex form in the time domain) most research concerned with the properties of estimators in the presence of autocorrelation has been conducted with the frequency domain representation. For this reason a decision was made to investigate the possibility of estimating ARMA /ARMAX models in the frequency domain, and the result was the development of an Aitken Estimator in the Frequency Domain (A.F.D.).

The model to be estimated will be ARMA (p,q)

$$y(t) + \beta_1 y(t-1) + \ldots + \beta_p y(t-p) = \varepsilon(t) + \alpha_1 \varepsilon(t-1) + \ldots + \alpha_q \varepsilon(t-q) \qquad (2.24).$$

It is assumed that $\varepsilon(t)$ is n.i.d. $(0,\sigma^2)$.[4]

Taking the finite Fourier Transform of (2.24) we obtain[5]

$$\sum_{t=0}^{N-1} (y(t)+\beta_1 y(t-1)+\ldots+\beta_p y(t-p))\, e^{i\lambda_j t} = \sum_{t=0}^{N-1} (\varepsilon(t)+\alpha_1 \varepsilon(t-1)+\ldots+\alpha_q \varepsilon(t-q))e^{i\lambda_j t}$$

$$j = 0,\ldots,N-1 \qquad (2.25).$$

A compact form of (2.25) is

$$B(\lambda_j)Y(\lambda_j) = A(\lambda_j)\varepsilon(\lambda_j) \qquad j = 0,\ldots,(N-1) \qquad (2.26)$$

where $B(\lambda_j) = 1+\beta_1 e^{i\lambda_j}+\ldots+\beta_p e^{i\lambda_j p}$

---

4 $\varepsilon(t)$ is adopted in preference to e(t) because of the special meaning of the latter symbol in Fourier Transform theory.

5 The finite Fourier Transform of a series Z(t) (t=0,...,N-1) is $\sum_{t=0}^{N-1} Z(t)e^{i\lambda_j t}$ where $\lambda_j = \frac{2\pi j}{N}$ (j=0,...,N-1). For frequency λ_j this will be designated as $Z(\lambda_j)$ but in most of what follows the subscript j will be omitted. The convention of denoting the periodogram of Z(t) as I_{ZZ} will also be followed and the j'th periodogram ordinate is given by $I_{ZZ}(\lambda_j) = \frac{1}{2\pi N} Z(\lambda_j)\bar{Z}(\lambda_j)$ where the bar denotes complex conjugation. The symbol is also used in Section 2.5.1 to distinguish parameter sets but it is unlikely that there will be any confusion.

$$Y(\lambda_j) = \sum_{t=0}^{N-1} Y(t)e^{i\lambda_j t}$$

$$A(\lambda_j) = 1 + \alpha_1 e^{i\lambda_j} + \ldots + \alpha_q e^{i\lambda_j q}$$

$$\varepsilon(\lambda_j) = \sum_{t=0}^{N-1} \varepsilon(t)e^{i\lambda_j t}.$$

We take as the starting point the function minimized in (2.16) i.e. $S=u^T\Omega^{-1}u$. Grenander and Szebö [34] have shown that

$$\operatorname*{plim}_{N\to\infty} \quad S = \tilde{S} = \sum_{j=0}^{N-1} \frac{I_{uu}(\lambda)}{f_{uu}(\lambda)} \tag{2.27}$$

where $I_{uu}(\lambda)$, $f_{uu}(\lambda)$ are respectively the periodogram and spectral density functions of the u(t) process.[6]

It is convenient at this stage to list a number of relations that will be manipulated in later analysis. The first two are definitional.

$$I_{uu}(\lambda) = \frac{1}{2\pi N} u(\lambda)\bar{u}(\lambda) \tag{2.28a}$$

$$|A(\lambda)|^2 = A(\lambda)\,\bar{A}(\lambda) \tag{2.28b}$$

From linear filtering theory we also have

$$u(\lambda) = Y(\lambda)B(\lambda) = A(\lambda)\varepsilon(\lambda) \tag{2.28c}$$

$$\varepsilon(\lambda) = Y(\lambda)B(\lambda)/A(\lambda) \tag{2.28d}$$

$$f_{uu}(\lambda) = \frac{\sigma^2}{2\pi} |A(\lambda)|^2 \tag{2.28e}.$$

Substituting for $I_{uu}(\lambda)$ from (2.28a) (and using (2.28c)) (2.27) is reduced to

$$\tilde{S} = \frac{1}{2\pi N} \sum_{j=0}^{N-1} \frac{\varepsilon(\lambda)\bar{\varepsilon}(\lambda)|A(\lambda)|^2}{f_{uu}(\lambda)} \tag{2.29}.$$

A further substitution of (2.28e) into this expression leaves

$$\tilde{S} = \frac{1}{2\pi N} \sum_{j=0}^{N-1} \frac{2\pi}{\sigma^2}\, \varepsilon(\lambda)\bar{\varepsilon}(\lambda) \tag{2.30}.$$

---

6 A proof may be found in Fishman [26].

Defining the $(p+q) \times N$ matrix Z and the $(N \times 1)$ vector W with elements

$$Z(k,j+1) = \frac{\partial \varepsilon(\lambda_j)}{\partial \theta_k} \qquad j = 0,\ldots,(N-1) \quad k = 1,\ldots,(p+q)$$

$$W(j+1) = \varepsilon(\lambda_j) \qquad j = 0,\ldots,(N-1)$$

the Gauss-Newton iterative sequence for the minimization of (2.30) will be

$$\theta^{(n)} - \theta^{(n-1)} = -(Z\bar{Z}^T)^{-1}\bar{Z}W \tag{2.31}$$

The component derivatives needed to construct Z are (using the form for $\varepsilon(\lambda)$ in (2.28d))

$$\frac{\partial \varepsilon(\lambda)}{\partial \alpha_k} = -\frac{B(\lambda)Y(\lambda)e^{i\lambda k}}{A^2(\lambda)} \qquad k = 1,\ldots,q$$

$$\frac{\partial \varepsilon(\lambda)}{\partial \beta_k} = \frac{Y(\lambda)e^{i\lambda k}}{A(\lambda)} \qquad k = 1,\ldots,p \tag{2.32}$$

Partitioning $Z\bar{Z}^T$ into $\begin{bmatrix} R & Q^T \\ Q & C \end{bmatrix}$ with typical elements of each block

$$R(k,\ell) = \sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)|B(\lambda)|^2 e^{i(k-\ell)\lambda}}{(|A(\lambda)|^2)^2} \qquad k,\ell = 1,\ldots,q$$

$$Q(k,\ell) = \sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)\bar{B}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2\bar{A}(\lambda)} \qquad k = 1,\ldots,p \quad \ell = 1,\ldots,q \tag{2.33}$$

$$C(k,\ell) = \sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad k,\ell = 1,\ldots,p$$

Similarly partitioning $\bar{Z}W$ into $\begin{bmatrix} D \\ F \end{bmatrix}$ we also have

$$D(k) = -\sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)\epsilon^{-i\lambda k}|B(\lambda)|^2}{|A(\lambda)|^2 \bar{A}(\lambda)} \qquad k = 1,\ldots,q$$

$$F(k) = \sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)e^{-ik} B(\lambda)}{|A(\lambda)|^2} \qquad \ell = 1,\ldots,p \qquad (2.34).$$

As an example of the derivation take $Q(k,\ell)$. This was deduced by first multiplying $\frac{\partial \epsilon(\lambda)}{\partial \beta_k}$ by $\frac{\partial \bar{\epsilon}(\lambda)}{\partial \alpha_\ell}$ and then by multiplying the resultant expression by $\frac{1}{2\pi N}$ thereby transforming $Y(\lambda)\bar{Y}(\lambda)$ to $I_{yy}(\lambda)$. (This scaling factor would not be required for estimation as it is common to both $Z\bar{Z}^T$ and $\bar{Z}W$ and therefore cancels, but it is of importance in evaluating S and the covariance matrix.)

To derive estimates of the residual variance and the covariance matrix we consider the log likelihood of (2.9)

$$\log L = K - \frac{N}{2}\log \sigma^2 - \frac{\pi}{\sigma^2}\sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)|B(\lambda)|^2}{|A(\lambda)|^2} \qquad (2.35).$$

At the maximum an estimate of the residual variance is obtained from $\frac{\partial \log L}{\partial \sigma^2} = 0$ which solves to yield

$$\hat{\sigma}^2 = \frac{2\pi}{N}\sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)|B(\lambda)|^2}{|A(\lambda)|^2} \qquad (2.36),$$

and an estimate of the covariance matrix will be

$$V_{\hat{\theta}\hat{\theta}} = (2\pi)^{-1}\hat{\sigma}^2(Z\bar{Z}^T)^{-1} \qquad (2.37).$$

When engaged in empirical research evaluation of (2.37) will be the easiest way of acquiring an estimate of the covariance matrix but under some circumstances the asymptotic covariance matrix may be of interest. An analytic form for this will be given in the following paragraphs.

The classical asymptotic theory of maximum likelihood estimators states that $\sqrt{N}(\hat{\theta}-\theta)$ has a limiting multinormal distribution with zero mean and covariance matrix I^{-1} where

$$I_{ij} = \lim_{N\to\infty} \frac{1}{N} E\left(- \frac{\partial^2 L}{\partial\theta_i \partial\theta_j} \right) \qquad (2.38).$$

Now the expression in brackets is proportional to the weighting matrix of the Newton-Raphson algorithm and it is shown in Appendix 3 that as $N\to\infty$ the expected value of the Gauss-Newton weighting matrix converges to that of the Newton-Raphson matrix enabling (2.38) to be written as

$$I = \lim_{N\to\infty} \frac{\sigma^2}{2\pi} \left(\frac{1}{N} E(Z\bar{Z}^T) \right)^{-1} \qquad (2.39).$$

Consider $\lim_{N\to\infty} \frac{1}{N} E(Z\bar{Z}^T)$ element by element. To illustrate the general argument we will concentrate on the sub-matrix R.

$$R(k,\ell) = \sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)|B(\lambda)|^2 e^{i(k-\ell)\lambda}}{(|A(\lambda)|^2)^2} \qquad k,\ell = 1,\ldots,q \qquad (2.40)$$

Exploiting the property $\lim_{N\to\infty} E(I_{yy}(\lambda)) = f_{yy}(\lambda)$ and the equality $f_{yy}(\lambda) = \frac{\sigma^2|A(\lambda)|^2}{2\pi|B(\lambda)|^2}$, (2.40) simplifies to

$$\lim_{N\to\infty} \frac{2\pi}{\sigma^2} E\left(\frac{1}{N} R(k,\ell) \right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \, d\lambda \qquad (2.41).$$

By repeating the above operations I may be reduced to the following asymptotic expression:-

$$I = \begin{bmatrix} I_1 & I_2^T \\ I_2 & I_3 \end{bmatrix}$$

$$I_1(k,\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(k-\ell)\lambda}}{|A(\lambda)|^2}\, d\lambda \qquad k,\ell = 1,\ldots,q$$

$$I_2(k,\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(k-\ell)\lambda}}{\bar{A}(\lambda)B(\lambda)}\, d\lambda \qquad \begin{matrix} k = 1,\ldots,p \\ \ell = 1,\ldots,q \end{matrix} \qquad (2.42).$$

$$I_3(k,\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(k-\ell)\lambda}}{|B(\lambda)|^2}\, d\lambda \qquad k,\ell = 1,\ldots,p$$

Equation (2.42), by showing that the asymptotic covariance matrix is a function of the ARMA parameters illustrates the use of the Fourier Transform as an expository and analytical device in econometrics. Although the result might also be obtained by the application of Z transform theory to ARMAX models, there are "spin-off" benefits to be had from working in the frequency domain viz. the theoretical spectra and response functions.[7]

An obvious application of (2.42) arises in testing the efficiency of the alternative estimators given earlier, as one may compare the empirical (Monte Carlo) variances with those calculated from (2.42) by a substitution of the known (β,α) parameters (or the means $\bar{\beta},\bar{\alpha}$ from all replications). This comparison enables an assessment of the efficiency of any estimator relative to the maximum likelihood estimator and provides a test on the sample size at which asymptotic theorems hold and in this role will appear in later chapters in the context of the Monte Carlo studies.

---

[7] The analogue of (2.42) for an ARMAX equation is derived in Appendix 4.

2.5 Alternative Time Domain Estimators

A number of other estimators were available in the literature either preceding or during the writing of this thesis and some justification for avoiding these methods is essential. The current section provides this for the estimators that are to be found in the "econometric" literature while Appendix 5 analyses those that appear in the "time series"literature. Also included in Chapter 5 is a detailed critique of a number of estimators that have been proposed for use in the frequency domain. As the estimators of Appendix 5 would be difficult to implement and were not specifically intended for use by economists such a division is appropriate.

2.5.1. e* and the S function

The three alternative estimators considered in the following sections differ according to the assumptions made concerning e* so that some order to the discussion will be achieved by presenting an account of the influence of e* upon the estimation of the parameters $[\delta,\alpha]$ and by establishing criteria which will enable a judgement of the magnitude of this effect. As the influence of e* is transmitted via the e series (equation (2.13)) to the sum of squares function (equation (2.10)) which discriminates between alternative sets of parameters, assessing the influence of e* upon e is equivalent to determining its impact upon S.

Consider two trial parameter sets $\underline{\theta} = (\underline{\delta}, \underline{\alpha}, \underline{e}^*)$ and $\bar{\theta} = (\bar{\delta}, \bar{\alpha}, \bar{e}^*)$ with the corresponding e vectors being $\underline{e}$ and $\bar{e}$ and the S functions $\underline{S}$ and $\bar{S}$. The fundamental relation for the analysis is

$$e(t) = u(t) - P(L)e(t) \qquad (2.43)$$

where $P(L) = \alpha_1 L + \ldots + \alpha_q L^q$ i.e. the M.A. polynomial with the normalized element deleted. Corresponding to the two parameter sets we have

$$\underline{e}(t) = \underline{u}(t) - \underline{P}(L)\underline{e}(t) \tag{2.44}$$

$$\bar{e}(t) = \bar{u}(t) - \bar{P}(L)\bar{e}(t) \tag{2.45}.$$

Imposing the restriction that $\underline{\theta}$ and $\bar{\theta}$ differ only in the values assumed for $\underline{e}^*$ and $\bar{e}^*$ (i.e. $\underline{\delta} = \bar{\delta}$, $\underline{\alpha} = \bar{\alpha}$) means that $\underline{u}(t) = \bar{u}(t)$ and $\underline{P}(L) = \bar{P}(L)$ so that substracting (2.45) from (2.44) we obtain

$$\underline{e}(t) - \bar{e}(t) = -\bar{P}(L)(\underline{e}(t) - \bar{e}(t)) \tag{2.46},$$

or

$$d(t)(1+\bar{P}(L)) = 0 \tag{2.47}$$

where

$$d(t) = \underline{e}(t) - \bar{e}(t).$$

Now $\bar{A}(L) = 1+\bar{P}(L)$ (where $\bar{A}(L)$ is the M.A. polynomial) so that (2.47) becomes

$$d(t)\bar{A}(L) = 0 \tag{2.48}.$$

To illustrate the use of (2.48) let $\underline{e}^*$ be the M.L. estimate of e^*, and let $\bar{e}^* = 0$. Then equation (2.48) allows the evaluation of the discrepancy between the M.L. estimate of $e(t)$ and the estimate obtained under the assumption that $e^* = 0$. In general the solution to (2.48) is a function of the roots of the characteristic equation and the initial conditions $d(-1),\ldots,d(-q)$: the roots being the inverse of those associated with $\bar{A}(L) = 0$ and therefore lying within the unit circle if Condition 7 of Chapter 1 is imposed upon $\bar{A}(L)$. Therefore $\lim_{N\to\infty} d(t) = 0$ i.e. the discrepancy will vanish as the sample size increases, so that regardless of the initial conditions or the size of the roots (subject to the satisfaction of Condition 7) the discrepancy between the M.L. and

any other estimate of e(t) will approach zero as the sample size tends to infinity - it is the rapidity of convergence which is determined by the size of roots and initial conditions.

In general, for any $\bar{e}^*$ (not necessarily zero) $\underset{N\to\infty}{\text{Lim}}\ \bar{S} = \underline{S}$ meaning that the extent to which the use of $\bar{S}$ in place of $\underline{S}$ will bias parameter estimates away from their M.L. magnitude will be a function of three factors. The bias will be larger

(i) The greater the magnitude of $d(-1),\ldots,d(-q)$.

(ii) The closer to unity is the modulus of the greatest root of the M.A. polynomial.

(iii) The smaller the sample size.

2.5.2. Estimates of e* by Concentration of the Likelihood

Further progress in investigating the effects of various assumptions about e* may be made by employing a technique frequently employed in the Full Information - Maximum Likelihood approach to the estimation of parameters in simultaneous equations - that of concentrating the likelihood with respect to a set of "nuisance" parameters. Adopting this course of action S will be concentrated with respect to e*.

Equation (2.14) presents the first derivative of e with respect to the j'th element of e* (i.e. e^*_j) and this generalizes to

$$\frac{\partial e}{\partial e^*} = -M^{-1}M^* \qquad (2.49).$$

Using (2.10) and (2.49) we have

$$\frac{\partial S}{\partial e^*} = -2(M^{-1}M^*)^T e \qquad (2.50).$$

Substituting for e from (2.13), (2.50) is reduced to

$$\frac{\partial S}{\partial e^*} = -2(M^{-1}M^*)^T M^{-1}(u - M^*e^*) \qquad (2.51).$$

The M.L. estimate of e^* is obtained by equating (2.51) to zero. Denoting (M^TM) by V this will be

$$\hat{e}^* = (M^{*T} V^{-1} M^*)^{-1} M^{*T} V^{-1} u \qquad (2.52).$$

The S function is now concentrated by replacing e^* in (2.13) by $\hat{e}^*$ from (2.52)

$$\therefore e = M^{-1}(u - M^*(M^{*T} V^{-1} M^*)^{-1} M^{*T} V^{-1} u) \qquad (2.53),$$

or if

$$\underset{\sim}{e} = M^{-1}u, \quad K = M^{-1}M^*$$

$$e = [I - K(K^TK)^{-1} K^T]\underset{\sim}{e} \qquad (2.54).$$

By initializing e^* at zero in (2.13) it is found that $\underset{\sim}{e}$ is the estimate of e resulting from the assumption that all pre-period values of e are zero, so that the M.L. estimate of e is constructed by applying a correction matrix to the $\underset{\sim}{e}$ vector.

At first glance concentration is attractive (computationally) as in many applications q may be quite high (e.g. in a system of equations or when dealing with weekly or monthly data) and this scheme, by separating "nuisance" from "fundamental" coefficients, considerably reduces the size of all inversions to be performed, thereby lessening the chance of rounding errors accumulating to dangerous levels. However the rise in the frequency of inversions (K^TK principally) occasioned by the use of a search routine to find the "optimal" step length along the Gauss-Newton (Newton-Raphson) direction and/or reliance on numerical evaluation of

derivatives may result in greater computational time being expended.[8] To overcome this three methods were experimented with:-

(i) A Steepest-Descent algorithm (which is based on only the first derivatives) was adopted for a few iterations in the hope that the algorithm would then be near the maximum and only a few iterations using the concentrated likelihood would be required. Unfortunately, as most users of Steepest-Descent would know, this solution was unreliable owing to slow convergence.

(ii) The assumption that $e^* = 0$ was made for three iterations and then concentration was applied. As the estimates were by then in the vicinity of the maximum it was important to evaluate S accurately. This suggestion was satisfactory for the Monte Carlo experiments, in which five iterations was the average number required for convergence, but not for the empirical examples which averaged around ten so that setting the switch point at three was likely to result in an excessive usage of computer time. One possibility (not explored) would be to make the switch on the basis of changes in the S function.

(iii) All derivatives were computed under the assumption that $e^* = 0$ but the S function was evaluated by concentration. This yielded M.L. estimates with large computational savings. Two reasons may be advanced for this. Firstly, differentiating (2.48) with respect

---

[8] For the Newton-Raphson version five function evaluations were required to compute one cross derivative so that in a five parameter model 75 function evaluations were required to construct all cross derivatives. This was a major factor in explaining the inordinate amount of computer time used by a version of the system of equations estimator discussed in Chapter 5.

to β_k (say) indicates that the derivatives formed under the assumption $e^* = 0$ converge to their true values. Secondly, a heuristic justification lies in the search routine which discriminates between parameter sets on the basis of the correct direction and only a rough indication of the order of magnitude of the step length, and both quantities are not greatly affected by the insertion of zeroes for e^*.

2.5.3 The Box-Jenkins-Watts-Bacon Solution

Box, Jenkins and Bacon [10] and Jenkins and Watts [55] seem to have been the first to suggest estimation under the assumption that $e^* = 0$. The dual justification advanced being that this is an unbiassed estimate and that as the sample size increases the initial values cease to be important. Section 2.5.1. has shown that the sample size required for the transient to die out is a function of the M.A. roots and a recognition of this points to the conclusion that this solution cannot have universal applicability. However as a first approximation it is useful.

2.5.4. The Box and Jenkins Solution

Box and Jenkins have provided a scheme to generate M.L. estimates of e^*. Essentially this involves concentration but differs from the analysis of 2.5.2. in the function concentrated. These authors distinguish between $S = e^T e$ (referred to as the sum of squares conditional upon estimates of e^*) and $\tilde{S} = e^T e + e^{*T} e^*$ (referred to as the unconditional sum of squares). Previously S has been concentrated and Box and Jenkins' contribution is to concentrate $\tilde{S}$.

By reworking 2.5.2. it may be shown (see Appendix 6) that (2.52) becomes

$$\hat{e}^* = (I_q + M^{*T} V^{-1} M^*)^{-1} M^{*T} V^{-1} u \qquad (2.55),$$

so that

$$e = [I - K(I_q + K^T K)^{-1} K^T] \underset{\sim}{e} \qquad (2.56).$$

Therefore the inclusion of e* into the function to be minimized will influence the estimates of e and e*. For the special case q=1 Appendix 6 proves that the two estimates of e* (denoted as $\underline{e}^*$ when obtained from a concentration of S and $\tilde{e}^*$ from $\tilde{S}$) are as $N \to \infty$

$$\underline{e}^* = \frac{u(0) - \alpha_1 u(1) + \alpha_1^2 u(2) - \alpha_1^3 u(3)\ldots}{\alpha_1}$$

$$\tilde{e}^* = \alpha_1(u(0) - \alpha_1 u(1) + \alpha_1^2 u(2) - \alpha_1^3 u(3).....)$$

or defining $u^* = u(0) - \alpha_1 u(1) + ...$

$$\underline{e}^* = u^* / \alpha_1 \qquad (2.57)$$

$$\tilde{e}^* = \alpha_1 u^* \qquad (2.58)^9.$$

From (2.57) and (2.58) we see that concentration of the two functions will yield radically different estimates of e* if α_1 is close to zero. Clearly however $\underset{\alpha_1 \to 1}{\text{Lim}}\ \underline{e}^* = \tilde{e}^*$ with the consequence that large differences appear between the two estimates only for low α_1 - the situation in which the transient effect dies out rapidly. Although analytical proof is not possible for all orders it is likely that unless the sample is very small there will be only a limited margin of choice between the

---

9 In Chapter 6 it is shown that u* corresponds to the adaptive backward forecast of u(-1) at time 0 as it is formed by an exponential smoothing of all past u(t) values. As this is the minimum mean square error forecast for a first order M.A. there is some ground for adopting a forecasting approach in the determination of e* (see Appendix 6).

estimators, but on balance one suspects that minimization of $\tilde{S}$ is preferable. This modification may be incorporated easily into the Phillips estimator by merely redefining S or by including the unit matrix in the equation defining concentrated estimates of e*. Chapter 4 will be concerned with differences between the two estimators when applied to simulated data so that some decision on the relative merits of S and $\tilde{S}$ may be made then.

2.5.5. The Klein-Dhrymes-Steiglitz Solution

Dhrymes, Klein and Steiglitz [22] have proposed an estimator rooted in previous papers by Steiglitz and McBride [100] and Steiglitz and Rogers [101]. The estimator differs from Phillips' in that the derivatives used in the Gauss-Newton algorithm are constructed by prefiltering the data.

The model to be estimated is (2.3) for which an equivalent form is

$$\frac{B(L)}{A(L)}\, y(t) = \frac{C(L)}{A(L)}\, x(t) + e(t) \qquad (2.59).$$

It is possible to give (2.59) a matrix representation as follows.

(i) An alternative expression of (2.59) is

$$w(t) = z(t) + e(t) \qquad (2.60a).$$

Obviously this implies that $w(t)A(L) = y(t)B(L)$, $z(t)A(L) = x(t)C(L)$.

(ii) Section 2.3.1 contains the derivation of the matrix form of the Phillips estimator. For the special case of an ARMA model this is (from (2.12))

$$e = M^{-1}(y - Y\beta - M^{*}e^{*}) \qquad (2.60b).$$

Now (2.60b) involves the construction of $e(t)A(L) = y(t)B(L)$ and a comparison of this with $w(t)A(L) = y(t)B(L)$ etc. clearly indicates that (2.60a) may be rewritten as

$$M^{-1}(y - Y\beta - M^{*}w^{*}) = M^{-1}(X\gamma - M^{*}z^{*}) + e \qquad (2.61)$$

where, analogously to e*,w* and z* are pre-period values.

Using (2.13) (2.61) reduces to

$$M^{-1} M^* w^* = M^{-1} M^* z^* + M^{-1} M^* e^* \qquad (2.62).$$

The Dhrymes Klein-Steiglitz estimator attempts to minimize the sum of squares associated with (2.61) by a Gauss-Newton algorithm. To generate initial estimates of w*, z* these authors comment that a "convenient initial condition ..... is w* = z* = 0". From (2.62) and the form of M* it is apparent that this is equivalent to initializing e* at zero i.e. this estimator is identical to that of Section (2.5.3) and as such shares the disadvantages sketched there for that estimator.

In passing we acknowledge the contribution of Maddala and Rao [70] who (in the context of a rational distributed lag model wherein the equality (2.62) would become M*z* + M*e* = 0) suggest the inclusion of z* in the parameter set to be estimated. As this is interchangeable with estimating e* it will possess the same sampling distribution as the Phillips estimator given earlier.

2.6 Computational Considerations

2.6.1 Starting Values

The selection of starting values is controversial. Phillips (and many other authors) favour consistent estimates and he provides algorithms to calculate these for δ and α. He also suggests that the unbiassed estimate of e* (zero) be used as an initial value for that parameter. At an early stage of research it was found (in Monte Carlo simulations) that the following scheme yielded the same parameters (to three decimal places) and the same likelihood (to about six decimal places) as that given by Phillips.

(i) Estimate δ by O.L.S.

(ii) Assume α, $e^* = 0$.

(iii) Iterate with (i) and (ii) as starting values.

Owing to the simplicity of (i) - (iii) it was adopted in the simulation studies with the proviso that if significant biases were found the experiment was re-run with the true parameters as the new starting values. Such checks, although unprofitable with the final algorithm used, were necessary with earlier versions. The general indication given by the Monte Carlo experiments was that the Gauss-Newton algorithm would converge from a wide variety of initial estimates: a feature that may justify the selection of quite arbitrary starting values (subject of course to the root restrictions being satisfied).

When fitting models to actual time series the above scheme was used unless there was some desire to reduce the computer time required for convergence. Then it was always better to obtain initial estimates that were likely to be close to the true values.[10] A strategy frequently availed of when there was no à priori information available concerning α was to generate initial estimates as follows:

Let $\hat{\gamma}(j)$ be the j'th covariance of the estimated residual vector $\hat{u}$.

Let $f_j(\alpha,\sigma^2)$ be the j'th theoretical covariance of u(t).

Then there are a system of (q+1) equations to be solved (exploiting

---

[10] In this connection it is worth noting the warning given by Freudenstein and Roth [27] that "The size of the domain of convergence depends upon the system of equations. For real algebraic equations, generally the size of the domain of convergence is inversely related to the degree and number of equations." There was a lack of convergence in one of the applications of Chapter 5 which seemed a consequence of the very large number of parameters present.

the zeroes in $f_j(\alpha,\sigma^2)$)

$$\hat{\gamma}(j) = f_j(\alpha,\sigma^2) \qquad j = 0,\ldots,q \qquad (2.63).$$

(2.63) constitutes a set of (q+1) equations that must be solved for the q parameters of α and σ^2. As f ( ) will be a non-linear function of α, σ^2 a Gauss-Newton algorithm was employed to produce estimates $\hat{\alpha}$, $\hat{\sigma}^2$ - the objective function to be minimized being

$$S = \sum_{j=0}^{q} (\hat{\gamma}(j) - f_j(\alpha,\sigma^2))^2 \qquad (2.64)^{11}.$$

Heuristically the algorithm attempts to match theoretical and empirical covariances as closely as possible. A perfect fit would be indicated by S = 0 but this is unlikely and in practice S varied from values of the order of 10^{-10} to 10^{-35} so that the method proved capable of providing reasonable starting values with minimal computing expense (1 second of C.P.U. time) and because of its flexibility found frequent use.

2.6.2. Constrained Estimation

A problem that may arise during the iterative sequence which may have unpredictable consequences is for the roots of $\hat{B}(L) = \hat{A}(L) = \hat{C}(L) = 0$ to violate Condition 7 of Chapter 1 and to guard against this it was

---

11 To give an example consider the second order M.A. $(1+\alpha_1 L+\alpha_2 L^2)e(t)$. The covariances $\hat{\gamma}(0)$, $\hat{\gamma}(1)$, $\hat{\gamma}(2)$ are calculated from the residuals, and $f_j(\alpha,\sigma^2)$ will be

$$f_0(\alpha,\sigma^2) = (1 + \alpha_1^2 + \alpha_2^2)\sigma^2$$
$$f_1(\alpha,\sigma^2) = (\alpha_1 + \alpha_1\alpha_2)\sigma^2$$
$$f_2(\alpha,\sigma^2) = \alpha_2\sigma^2$$

which yields 3 equations to be solved in the three unknowns α_1, α_2, σ^2.

necessary to enforce this restriction during the iterations.

Enforcement was achieved by the use of a "penalty function" concept in that if the modulus of any of the roots, when tested, was less than unity, S was set at 1.0×10^{35} so that these parameters were rejected and the search would be continued elsewhere.[12] For low order polynomials it is possible to solve for the roots directly but with higher order cases e.g. $(1 + \alpha_1 L + \alpha_4 L^4 + \alpha_5 L^5)$ most polynomial routines available to the author performed badly so that a simpler method of testing was required.

The Routh-Hurwicz theorem (Baumol [7 p.257]) can be used for this purpose.

Theorem: The zeroes of the polynomial equation

$$a_0 + a_1 x + \ldots\ldots + a_n x^n = 0$$

will all lie outside the unit circle if and only if the following n determinants are positive.

$$\begin{vmatrix} a_0 & \cdot & a_n \\ \cdots & \cdot & \cdots \\ a_n & \cdot & a_0 \end{vmatrix} \quad \begin{vmatrix} a_0 & 0 & \cdot & a_n & a_{n-1} \\ a_1 & a_0 & \cdot & a_0 & a_n \\ \cdots & \cdots & \cdot & \cdots & \cdots \\ a_n & 0 & \cdot & a_0 & a_1 \\ a_{n-1} & a_n & \cdot & 0 & a_0 \end{vmatrix} \quad \begin{vmatrix} a_0 & 0\ldots\ldots 0 & \cdot & a_n & a_{n-1}\ldots\ldots a_1 \\ a_1 & a_0\ldots\ldots 0 & \cdot & 0 & a_n\ldots\ldots a_2 \\ \cdots & \cdots & \cdot & \cdots & \cdots \\ a_{n-1} & a_{n-2}\ldots a_0 & & 0 & 0\ldots\ldots a_n \\ \cdots & \cdots & & \cdots & \cdots \\ a_n & 0\ldots\ldots 0 & \cdot & a_0 & a_1\ldots\ldots a_{n-1} \\ a_{n-1} & a_n\ldots\ldots 0 & \cdot & 0 & a_0\ldots\ldots a_{n-2} \\ \cdots & \cdots & \cdot & \cdots & \cdots \\ a_1 & a_2\ldots\ldots a_n & \cdot & 0 & 0\ldots\ldots a_0 \end{vmatrix}$$

As the condition stands it is unwieldy. However it may be shown that there is an equivalent form which is a good deal simpler to implement.[13] Define the polynomial as above and set up the elements of the matrix A

12 Penalizing S was found to be more satisfactory than the alternative of reducing the step length until the root restriction was satisfied.

13 See Gantmacher [29].

with the initialization

$$A(0,j) = a_j \qquad j = 0,\ldots,n$$

and with all other elements defined by the following sequence of operations,

$$A(i,j) = A(i-1,0)A(i-1,j) - A(i-1,n-i+1)A(i-1,n-j-i+1)$$

$$i=1,\ldots,n$$
$$j=0,\ldots,(n-i) \qquad (2.65).$$

The determinantal conditions are then replaced by the condition that the roots of the polynomial equation will all be less than unity if and only if $A(i,0) > 0 \; \forall \; i = 0,\ldots,n$.

2.6.3 Improved Convergence

There are two aspects of convergence dealt with under this heading.

(i) Techniques to guard against a possible lack of convergence.

(ii) Techniques to accelerate convergence.

(i) Lack of Convergence

It is well known (Crockett and Chernoff [20]) that the weighting matrix must be positive definite if the sequence of iterations is to converge to a minimum of the function. With statistical problems this is assured at the minimum of the sum of squares function because of the equivalence of the weighting matrix and the covariance matrix of the estimates. However the region in which this weighting matrix remains positive definite may be quite restricted, so that if starting points are far from the minimum there may well be a lack of convergence.

In most applications dealt with in this thesis the matrix remained positive definite even for very poor starting values e.g. one equation fitted reduced the sum of squares from 810 to 0.02 in fifteen iterations, so that a lack of convergence was rare. Nevertheless a safeguard was

incorporated into all programs involving the addition of a positive scalar λ to the diagonal elements of the weighting matrix if any eigenvalue of this matrix was negative. As Shanno has shown [99] if λ is selected as greater than or equal to the negative of the smallest eigenvalue of this matrix the sequence of iterations constructed from the augmented matrix will converge to a minimum. In the scalar case it is easy to comprehend the logic of this step as the weighting matrix will collapse to a scalar and by ensuring that this is positive the Steepest-Descent direction is preserved, and as Akaike [3] has shown following this will eventually lead to the function minimum. Although Shanno furnishes a convergence proof there is no guarantee that the modified algorithm will not exhibit the slowness of Steepest-Descent. Nevertheless it is true that there will be a switch to a better algorithm once positive definiteness is established so that programs incorporating this feature are likely to be more efficient than pure Steepest-Descent versions.

An example will illustrate the technique. In Chapter 4 A.F.D. is applied to a series New Money and to test for the possibility of a local minimum multiple starting values were employed. For one of these parameter sets the weighting matrix was negative definite and Table 2.1 sets out the sequence of iterations for the Shanno scheme.

Table 2.1 Convergence of the Shanno Scheme

Iteration	Smallest Eigenvalue	α_1	α_4	S
0	-0.103×10^{5}	0.60	0.30	0.146×10^{5}
1	0.761×10^{4}	-0.17	-0.56	0.520×10^{4}
2	0.773×10^{4}	-0.19	-0.59	0.502×10^{4}
3	0.808×10^{4}	-0.19	-0.59	0.501×10^{4}

At a number of points in the thesis this modification enabled convergence so that it is worthwhile incorporating into non-linear algorithms.

(ii) Accelerated Convergence.

As well as the possibility of a lack of convergence there is the well known tendency of the Gauss-Newton algorithm to exhibit slow convergence if the quadratic approximation is a poor one, and to overcome this some restriction must be placed upon the step size.[14] This may be done by restricting each step to lie within a hypersphere of radius r to form the Lagrangian

$$H = S + \lambda\{(\theta^{(n)} - \theta^{(n-1)})^2 - r^2\} \tag{2.66}$$

where λ is the Lagrange multiplier and $\theta^{(n)}$ the estimate of θ at the n'th iteration.

Adopting the convention that all arrays not explicitly indexed are evaluated at the (n-1)'th iteration the following normal equations are obtained by minimizing $H(D = \frac{\partial e}{\partial \theta})$

$$\frac{\partial H}{\partial \theta} = 2(D^T e + D^T D(\theta^{(n)} - \theta^{(n-1)})) - 2\lambda(\theta^{(n)} - \theta^{(n-1)}) = 0 \tag{2.67}$$

$$\therefore \quad \theta^{(n)} - \theta^{(n-1)} = -(D^T D - \lambda I)^{-1} D^T e \tag{2.68}.$$

The constrained step of (2.68) is similar to that of the previous section and would be identically equal if λ was the minimum eigenvalue of $D^T D$.

---

[14] Slow convergence comes from overshooting the minimum necessitating a reversal of the step at the next iteration. An example would be Rosenbrock's function.

An alternative expression for λ was put forward by Levenberg [68]. From the first order condition $\frac{\partial H}{\partial \lambda} = 0$ we obtain $(\theta^{(n)} - \theta^{(n-1)})^2 = r^2$ which when substituted into (2.68) gives an expression for r^2.

$$r^2 = e^T D^T (D^T D - \lambda I)^{-2} D^T e \qquad (2.69)$$

Now $\lim_{\lambda \to \infty} (D^T D - \lambda I) \to \lambda I$ so that (2.68) becomes (with some re-arrangement)

$$\lambda^2 = e^T D^T D\, e / r^2 \qquad (2.70),$$

and

$$\lambda = \pm \sqrt{\lambda^2} \qquad (2.71).$$

Therefore by specifying the maximum allowable step length r it is possible to compute λ and thereby define a feasible parameter space at any iteration. As Marquardt has noted [73], the Levenberg parameter λ in ranging from zero to infinity causes the step length to shift from that of Gauss-Newton to Steepest-Descent.[15] Furthermore Marquardt's algorithm, which is based on this approach, has a long history of success and provides some justification for loading the diagonals of the weighting matrix in the above manner. In all applications r was set at unity and the negative value of λ from (2.71) was selected in order to preserve the Steepest-Descent direction.

2.7 Asymptotic Theory

Previous sections have concentrated upon the solution of the normal equations to yield estimates $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$. These estimates would be consistent, asymptotically normal and efficient if the disturbances u(t) were n.i.d. For ARMA/ARMAX models this independence condition is not present and it is natural to ask which properties still hold for

---

[15] Away from the minimum $|D| > > 0$ so that λ will be large whereas as the minimum is approached $|D| \to 0$, and therefore $\lambda \to 0$ i.e. the Gauss-Newton step becomes operative.

the estimates given in earlier sections.

Walker [114] and Whittle [116] have examined this in some detail for ARMA models and the following statements concerning the distribution of $\hat{\theta}$ (and the assumptions necessary for the proof) are taken directly from Walker.

ASSUMPTIONS

(i) $x(t)$ $(t=1,\ldots,N)$ is any linear process with finite fourth moment of the form

$$x(t) = \sum_{j=0}^{\infty} \alpha_j \, e(t-j)$$

where e(t) is i.i.d. $(0,\sigma^2)$, $0 < E\,(e^4(t)) < \infty$ and $\sum_{j=0}^{\infty} \alpha_j^2 < \infty$.

(ii) The estimates $\hat{\theta}_N$, $\hat{\sigma}_N^2$ are such that $L_N^*(\theta,\sigma^2)$ (the log likelihood) is an absolute maximum when $\theta = \hat{\theta}_N$, $\sigma^2 = \hat{\sigma}_N^2$.

(iii) The true values of the parameters lie in a region defined by $0 < \sigma^2 < \infty$, $\theta \in H$ where H is a bounded closed set contained in an open set S in k-dimensional Euclidean space. In particular the set H restricts all roots of $B(L) = A(L) = 0$ to lie absolutely outside the unit circle. This is not a severe restriction however as unit roots in the A.R. (the most likely case) will be prescribed rather than estimated.

(iv) Conditions 4 and 5 of Chapter 1 hold to assure identification.

(v) The frequency response function of the linear process $A(\lambda)$ and its inverse $A^{-1}(\lambda)$ are continuous functions of λ for $-\pi \leq \lambda \leq \pi$.

ASYMPTOTIC THEOREMS

Under the five assumptions above Walker proves the following theorems (p.365)

(A) $\hat{\theta}_N$ is a consistent estimate of θ i.e. $\underset{N\to\infty}{\text{plim}}\ \hat{\theta}_N = \theta$,

(B) $\sqrt{N}(\hat{\theta}_N - \theta)$ has, as $N\to\infty$, a limiting distribution which is multinormal with zero mean and covariance matrix equal to the inverse of the matrix $V = (v_{ij})$, where

$$v_{ij} = \underset{N\to\infty}{\text{Lim}}\ N^{-1}\, E\left(- \frac{\partial^2 L_N^*}{\partial\theta_i\, \partial\theta_j} \right).$$

(C) $\underset{N\to\infty}{\text{plim}}\ \hat{\sigma}_N^2 = \sigma^2$,

(D) $\sqrt{N}\ (\hat{\sigma}_N^2 - \sigma^2)$ has, as $N\to\infty$, a limiting normal distribution with mean zero and variance $2\sigma^4$, and in the limit $\sqrt{N}(\hat{\sigma}_N^2 - \sigma^2)$ and $\sqrt{N}(\hat{\theta}_N - \theta)$ are independently distributed.

A few comments upon the assumptions and the theorems are in order.

(i) Although the constraints imposed upon the parameter set and the x(t) series appear restrictive, in practice various transformations are available to construct series with these properties e.g. differencing and logarithms.

(ii) No proof is available if one of the parameters of θ is associated with an exogenous variable i.e. the model is ARMAX and not ARMA. This extension has proven to be difficult. Pierce has presented a solution [91] but there is some doubt concerning his proof of consistency and as he recognizes in the following quote there is an unresolved issue in the uniqueness of any estimate.

"One would expect that this solution would also be the unique consistent estimate of γ, but no proof of this is given here" [91 p.306].

As many of the properties invoked for his estimator rely on uniqueness this is a major stumbling block to an acceptance of his claim. Clearly this is one of the unsettled questions that must be the subject of future investigation.

(iii) The disturbances need only be i.i.d. and not n.i.d.

(iv) The estimates $\hat{\theta}, \hat{\sigma}^2$ must be such as to globally maximize the likelihood.

It is (iv) that is of greatest interest to this section as any invocation of the asymptotic theorems certainly requires that this be satisfied. Further progress may be made in considering (iv) by turning it into a query: does the non-linear algorithm converge to a global maximum? This question has two distinct components:

(A) Do the estimates define a maximum to L*?

(B) Is this maximum global or local?

Essentially A is concerned with convergence to a turning point.[16] The safeguards outlined earlier in the Chapter (see 2.6), by ensuring that no step is taken unless L* increases, result in a termination of the algorithm only when L* has reached a maximum e.g. see Hartley [48] for a

---

[16] There is a related question of whether the necessary conditions $\frac{\partial L^*}{\partial \theta} = 0$ are sufficient. From elementary calculus it is clear that this is not so but there is protection against finding a maximum (if we seek a minimum) by the program searching only for the latter. Although a saddle point is possible the fact that if the global minimum has been reached the weighting matrix is positive definite allows discrimination between the two cases.

proof of this conjecture. Even if the safeguards were not effective a suitable combination of an always-convergent Steepest-Descent method and a Gauss-Newton (or Newton-Raphson) algorithm - the switching point being defined by negative definiteness of the weighting matrix - would always ensure that a maximum was attained.[17]

To answer (B) there is no non-linear algorithm currently available that would select a global rather than a local minimum. Although there are some search procedures which allow a certain level of (prespecified) probability to be assigned to the event these require such a large number of function evaluations that the cost is prohibitive except for the simplest of functions. That such multiple maxima may exist with ARMA/ARMAX models has been demonstrated by Struik [102] (in connection with Box and Jenkins airline data) and Steiglitz and Rogers [101] (with simulated data). In both reports the sample size was at least 100. Given the cost of searching the parameter space multiple maxima may best be isolated by adopting different starting values, but even then there is no certainty that all will be found.

2.8 Conclusion

Chapter 2 (and Appendix 6) have surveyed a number of estimators that have been proposed in the literature for the estimation of ARMA/ARMAX models and has opted for techniques that obtain estimates by iterative solution of the non-linear likelihood equations. However as has been emphasized in 2.7 little is known of the sampling properties

---

17 As mentioned in an earlier section a proof of convergence does not provide any evidence on the rapidity of convergence, so that Steepest-Descent - despite a sound theoretical basis - would not be applied unless very large amounts of computer time were available.

of such estimators (even in large samples) and it will be the task of Chapters 3 and 4 to provide some insight into these questions.

Appendix 3

Asymptotic Equivalence of the Gauss-Newton and Newton-Raphson Weighting Matrices

To investigate the asymptotic equivalence we proceed as follows. Assuming that θ is a $(M\times 1)$ vector the required first and second derivatives are:

$$S = e^T e \tag{A3.1}$$

$$\frac{\partial S}{\partial \theta_k} = 2\,\frac{\partial e^T}{\partial \theta_k}\,e \qquad k = 1,\ldots,M \tag{A3.2}$$

$$\frac{\partial^2 S}{\partial \theta_\ell \partial \theta_k} = 2\left(\frac{\partial^2 e}{\partial \theta_\ell \partial \theta_k}\right)^T e + 2\left(\frac{\partial \epsilon}{\partial \theta_k}\right)^T \left(\frac{\partial e}{\partial \theta_\ell}\right) \quad k,\ell=1,\ldots,M \tag{A3.3}$$

The Gauss-Newton and Newton-Raphson weighting matrices will be equal (up to a constant of proportionality 2) if

$$\left(\frac{\partial^2 e}{\partial \theta_\ell \partial \theta_k}\right)^T e = 0 \qquad k,\ell = 1,\ldots,M \tag{A3.4}$$

For the covariance matrix of A.F.D. this condition is

$$\lim_{N\to\infty} \frac{1}{N} E\left(\sum_{j=0}^{N-1} \frac{\partial^2 \bar{\epsilon}(\lambda)}{\partial \theta_\ell \partial \theta_k}\,\epsilon(\lambda)\right) = 0 \tag{A3.5}$$

Specializing to an ARMA (p,q) model (i.e. $\theta = \begin{bmatrix}\beta\\ \alpha\end{bmatrix}$) from (2.32) of the text the first and second derivatives of $\epsilon(\lambda)$ with respect to α and β are

$$\frac{\partial \epsilon(\lambda)}{\partial \alpha_k} = -\frac{B(\lambda)Y(\lambda)e^{i\lambda k}}{A^2(\lambda)} \qquad k=1,\ldots,q \tag{A3.6}$$

$$\frac{\partial \epsilon(\lambda)}{\partial \beta_k} = \frac{Y(\lambda)e^{i\lambda k}}{A(\lambda)} \qquad k=1,\ldots,p \tag{A3.7}$$

$$\frac{\partial^2 \epsilon(\lambda)}{\partial \alpha_\ell \partial \alpha_k} = \frac{B(\lambda)Y(\lambda)e^{i(k+\ell)\lambda}}{A^3(\lambda)} \qquad k,\ell=1,\ldots,q \tag{A3.8}$$

$$\frac{\partial^2 \varepsilon(\lambda)}{\partial \beta_\ell \partial \beta_k} = 0 \qquad k,\ell = 1,\ldots,p \qquad \text{(A3.9)},$$

$$\frac{\partial^2 \varepsilon(\lambda)}{\partial \alpha_\ell \partial \beta_k} = -\frac{Y(\lambda)e^{i(k+\ell)}}{A^2(\lambda)} \qquad \begin{matrix} k = 1,\ldots,p \\ \ell = 1,\ldots,q \end{matrix} \qquad \text{(A3.10)}.$$

Consider the limit of (A3.5) element by element.

$$\lim_{N\to\infty} \frac{1}{N} E\left[\sum_{j=0}^{N-1} \frac{\partial^2 \bar{\varepsilon}(\lambda)}{\partial \beta_\ell \partial \beta_k}\ \varepsilon(\lambda)\right] = 0 \qquad \text{(A3.11)}$$

$$\lim_{N\to\infty} \frac{1}{N} E\left[\sum_{j=0}^{N-1} \frac{\partial^2 \bar{\varepsilon}(\lambda)}{\partial \alpha_\ell \partial \beta_k}\ \varepsilon(\lambda)\right] = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{f_{yy}(\lambda)e^{-i(k+\ell)\lambda}B(\lambda)}{\bar{A}^2(\lambda)\,A(\lambda)}\,d\lambda \qquad \text{(A3.12)}$$

$$\lim_{N\to\infty} \frac{1}{N} E\left[\sum_{j=0}^{N-1} \frac{\partial^2 \bar{\varepsilon}(\lambda)}{\partial \alpha_\ell \partial \alpha_k}\ \varepsilon(\lambda)\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|B(\lambda)|^2 f_{yy}(\lambda)e^{-i(k+\ell)\lambda}}{|A(\lambda)|^2\,\bar{A}^2(\lambda)}\,d\lambda \qquad \text{(A3.13)}$$

With the substitution $f_{yy}(\lambda) = \dfrac{\sigma^2|A(\lambda)|^2}{2\pi|B(\lambda)|^2}$ (A3.12) and (A3.13) become

$$-\frac{\sigma^2}{(2\pi)^2}\int_{-\pi}^{\pi} \frac{e^{-i(k+\ell)\lambda}}{\bar{B}(\lambda)\bar{A}(\lambda)}\,d\lambda \qquad \text{(A3.14)}$$

$$\frac{\sigma^2}{(2\pi)^2}\int_{-\pi}^{\pi} \frac{e^{-i(k+\ell)\lambda}}{\bar{A}^2(\lambda)}\,d\lambda \qquad \text{(A3.15)}.$$

From the theory of residues the integrals in (A3.14) and (A3.15) converge to zero so that

$$\lim_{N\to\infty} \frac{1}{N} E\left[\frac{\partial \bar{\varepsilon}(\lambda)^T}{\partial \theta_k}\ \frac{\partial \varepsilon(\lambda)}{\partial \theta_\ell}\right] = \frac{1}{2} \lim_{N\to\infty} \frac{1}{N} E\left[\frac{\partial^2 S}{\partial \theta_\ell \partial \theta_k}\right] \sim$$

Furthermore $\dfrac{\partial^2 L}{\partial \theta_\ell \partial \theta_k} = 2\,\dfrac{\partial^2 S}{\partial \theta_\ell \partial \theta_k}$ so that the expected value of the Gauss-Newton weighting matrix is asymptotically the covariance matrix of $\sqrt{N}(\hat{\theta}-\theta)$.

Appendix 4

Derivation of the Asymptotic Covariance Matrix for ARMAX Estimators

The ARMAX model is

$$B(L)y(t) + C(L)x(t) = A(L)\varepsilon(t) \qquad (A4.1)$$

where $B(L)$ is of order p, $C(L)$ of order r and $A(L)$ of order q.

In Fourier Transform notation (A4.1) is

$$B(\lambda)\, Y(\lambda) + C(\lambda)\, X(\lambda) = A(\lambda)\, \varepsilon(\lambda) \qquad (A4.2).$$

The covariances will be formed from the following three sets of derivatives:

$$\frac{\partial \varepsilon(\lambda)}{\partial \beta_k} = \frac{Y(\lambda)e^{ik\lambda}}{A(\lambda)} \qquad k = 1,\ldots,p \qquad (A4.3)$$

$$\frac{\partial \varepsilon(\lambda)}{\partial \gamma_k} = \frac{X(\lambda)e^{ik\lambda}}{A(\lambda)} \qquad k = 0,\ldots,r \qquad (A4.4)$$

$$\frac{\partial \varepsilon(\lambda)}{\partial \alpha_k} = -\frac{\varepsilon(\lambda)e^{ik\lambda}}{A(\lambda)} \qquad k = 1,\ldots,q \qquad (A4.5).$$

Denoting H as the Gauss-Newton weighting matrix[1] the blocks of H are (where hats are omitted from the parameters)

$$H_{\beta\beta} = \sum_{j=0}^{N-1} \frac{I_{yy}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad k,\ell = 1,\ldots p \qquad (A4.6)$$

$$H_{\beta\gamma} = \sum_{j=0}^{N-1} \frac{I_{yx}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad \begin{matrix} k = 1,\ldots,p \\ \ell = 0,\ldots,r \end{matrix} \qquad (A4.7)$$

$$H_{\beta\alpha} = -\sum_{j=0}^{N-1} \frac{I_{y\varepsilon}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad \begin{matrix} k=1,\ldots,p \\ \ell= 1,\ldots,q \end{matrix} \qquad (A4.8)$$

$$H_{\gamma\gamma} = \sum_{j=0}^{N-1} \frac{I_{xx}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad k,\ell = 0,\ldots,r \qquad (A4.9)$$

---

[1] The argument of Appendix 3 extends to the ARMAX case.

$$H_{\gamma\alpha} = -\sum_{j=0}^{N-1} \frac{I_{x\varepsilon}(\lambda)e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad \begin{matrix} k = 0,\ldots,r \\ \ell = 1,\ldots,q \end{matrix} \qquad (A4.10)$$

$$H_{\alpha\alpha} = \sum_{j=0}^{N-1} \frac{I_{\varepsilon\varepsilon}(\lambda)\, e^{i(k-\ell)\lambda}}{|A(\lambda)|^2} \qquad k\ ,\ \ell = 1,\ldots,q \qquad (A4.11)$$

As in the text the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta}-\theta)$ is $V^{-1} = \lim_{N\to\infty} \frac{2\pi}{\sigma^2} E(\frac{1}{N} H)$ and by taking the limit we obtain the $(k\ \ell)$'th element of each block as

$$V_{\beta\beta} = \frac{1}{\sigma^2}\int_{-\pi}^{\pi} \frac{f_{yy}(\lambda)\, e^{i(k-\ell)\lambda}}{|A(\lambda)|^2}\, d\lambda \qquad k,\ell = 1,\ldots,p \qquad (A4.12)$$

$$V_{\beta\gamma} = \frac{1}{\sigma^2} \int_{-\pi}^{\pi} \frac{f_{yx}(\lambda)\, e^{i(k-\ell)\lambda}}{|A(\lambda)|^2}\, d\lambda \qquad \begin{matrix} k = 1,\ldots,p \\ \ell = 0,\ldots,r \end{matrix} \qquad (A4.13)$$

$$V_{\beta\alpha} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i(k-\ell)\lambda}}{\bar{A}(\lambda)B(\lambda)}\, d\lambda \qquad \begin{matrix} k = 1,\ldots,p \\ \ell = 1,\ldots,q \end{matrix} \qquad (A4.14)$$

$$V_{\gamma\gamma} = \frac{1}{\sigma^2} \int_{-\pi}^{\pi} \frac{f_{xx}(\lambda)\, e^{i(k-\ell)\lambda}}{|A(\lambda)|^2}\, d\lambda \qquad k,\ell = 0,\ldots,r \qquad (A4.15)$$

$$V_{\gamma\alpha} = 0 \qquad \begin{matrix} k = 0,\ldots,r \\ \ell = 1,\ldots,q \end{matrix} \qquad (A4.16)$$

$$V_{\alpha\alpha} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i(k-\ell)\lambda}}{|A(\lambda)|^2}\, d\lambda \qquad k,\ell = 1,\ldots,q \qquad (A4.17).$$

Equations (A4.12) - (A4.17) were used to evaluate the asymptotic covariances given in the Tables of Chapters 3 and 4.

Appendix 5

A Survey of Some Time Series Contributions to the Estimation of ARMAX Models

1. Methodology

This appendix provides a critique of a number of time series contributions to the estimation of ARMAX models. It is not meant to be an original discussion but to indicate the reasons for rejecting these estimators when faced with the need to apply such models to economic series. As such the three criteria to be applied in any judgement could not be regarded as having any theoretical basis but explain why these estimators play a very minor role in the thesis.

(i) Generalization

Under this designation we will explore the possibilities of generalizing algorithms presented by various authors for the estimation of a specific functional form to other forms. Such a generalization is imperative inasmuch as the range of model employed by econometricians is quite extensive, so that unless an estimator is applicable to all (or even most) of these it is of greatly diminished value. Two aspects of this are of particular importance:-

(a) Generalization of the same model to higher order lags than that considered by the original author e.g. an ARMA model estimator which has been tested upon ARMA (1,1) and ARMA (2,2) forms must be capable of estimating ARMA (5,5) equations as the latter may be essential for a parameterization of quarterly time series.

(b) Generalization to a different model but one of the same order i.e. there is a concern with the estimation of both ARMA and ARMAX forms. As econometrics has traditionally dealt with interrelationships between series such an extension is highly desirable.

(ii) Computational Simplicity and Relation to Numerical Methods

No objection is likely to be raised to the postulate of computational simplicity as a goal but a different (although no less important) aspect of computation is the relation of proposed techniques to the numerical methods field. A close relationship will make progress in the design of efficient computational algorithms a function of advances in numerical methods. As an example we cite the ever present danger of a lack of convergence to any final solution involved in the iterative schemes of Chapter 2 and this Appendix. By relating estimators to a field that is constantly attempting to overcome this problem the number of potentially successful applications will become larger. In any case both simple and efficient computational designs are more easily achieved if a number of alternatives can be explored.

(iii) Asymptotic Distributional Theory

Owing to the stochastic nature of the problem a knowledge of the distributional properties of the estimators becomes important. A minimal requirement might be the derivation of asymptotic sampling distributions. As a guiding principle we acknowledge that an estimator with known limiting properties has established a clear claim for adoption.

With the exception of (iii) the discussion of Chapter 2 has shown that the Phillips , A.T.D. and A.F.D. estimators satisfy these criteria so that there is a presupposition in favour of their general use.

2. Durbin's Estimator

Durbin [23], in one of the earliest contributions to the field, proposed an estimator along the following lines. Assume (for expository purposes) that the process to be modelled is a first order M.A. i.e. $y(t) = (1+\alpha L)e(t)$. If the invertibility condition is satisfied this may be converted to an infinite A.R.

$$y(t)(1+\alpha L+\alpha^2 L^2+\alpha^3 L^3+.....) = e(t) \qquad (A5.1).$$

Truncating at L^k, the parameters of the finite A.R. $(\beta_1,...,\beta_k)$ may be estimated by O.L.S. and by the theorems of Mann and Wald [72] $\sqrt{N}(\hat{\beta}_{OLS} - \beta)$ is asymptotically multinormal with zero mean and covariance matrix V. By exploiting the functional relationship between V and the M.A. parameter it is possible to derive a set of equations linear in α which, when solved, yield an estimate $\hat{\alpha}$.

How well does Durbin's estimator fare when judged by the criteria of Section 1?

(i) Generalization

Durbin in a later paper [24] has generalized the above methodology to ARMA but not ARMAX models.

(ii) Computational Ease

It is easily proven that the choice of k should be a function of the modulus of the largest root of the M.A. process so that if the Mann and Wald theorem is to hold (implying that the remainder $\sum_{j=k+1}^{\infty} \beta_j y(t-j)$ is small) k must be set according to the true

(unknown) roots of the M.A. polynomial. For an ARMA (0,1) model with $\alpha = 0.96$, $\beta_{60} = 0.3$ and this dictates the choice of at least a 60'th order A.R. before any of the Mann and Wald theorems will hold. Of course to estimate such high order A.R.'s is an impossibility given the collinearity characteristic of most economic data. Even if this obstacle were overcome inclusion of exogenous variables (each of which must be lagged an equal number of times as y(t)) may lead to the predicament of having to estimate very large numbers of parameters e.g. the above M.A. would yield 240 parameters if four exogenous variables are present. Computationally this is just not tenable.

(iii) Theoretical Properties

Hannan [44] has noted that it is difficult to prove efficiency of the estimator unless some assumption about $\lim_{N\to\infty} k/N$ is made. A proof might be constructed by relating k to the roots, so that given a parameter set k is a constant with the implication that $\lim_{N\to\infty} k/N = 0$, but this was not attempted. Another doubt concerning the efficiency of the estimator stems from the observation that the O.L.S. estimates are not fully efficient owing to the restrictions existing between the A.R. parameters e.g. for a first order M.A. $\beta_k = \alpha\beta_{k-1}$.

3. Walker's Estimator

Walker [112] [113] has derived a M.L. estimator based upon the asymptotic distribution of the sample serial correlations. Briefly, the distribution of k of these is considered, the likelihood is constructed,

and maximization of this is attempted. An iterative algorithm (corresponding to Gauss-Newton) is suggested as a suitable means of effecting this. Comments may be grouped under the previous criteria as follows:

(i) Generalization

Walker's articles apply to ARMA but not to ARMAX models. However there does not seem to be any objection to an extension to models linear in β and γ as the serial covariances will be functions of all parameters but it would be troublesome in generalizing the procedure to models non-linear in β and γ e.g. in Investment demand equations the rate of depreciation enters in a non-linear fashion.

(ii) Computational ease

A workable program may be distilled from Walker's article along the following lines. Let $r_j (j = 1,\ldots,k)$ be the sample serial correlation coefficients and ρ_j the corresponding population quantities and arrange ρ_j in a (k x 1) vector P. Define the (kx1) vector S with elements

$$
\begin{aligned}
S_\ell &= r_\ell && \ell = 1,\ldots,q \\
S_\ell &= \sum_{i=o}^{p} \beta_i \, r_{\ell-i} && \ell = (q+1),\ldots,(q+p) \qquad \text{(A5.2)} \\
S_\ell &= \sum_{i,j=o}^{p} \beta_i \beta_j \, r_{\ell-i-j} && \ell = (q+p+1),\ldots,k.
\end{aligned}
$$

Then Hannan has shown [47 p.40-41] that the joint distribution of $r_1, r_2, \ldots, r_k$ is asymptotically multinormal so that as $N \to \infty$ $\sqrt{N}(S-P)$ will be multinormal with mean 0 and covariance matrix W and this enables us to write the associated likelihood as

$$L_k = \frac{1}{2} k \log 2\pi - \frac{1}{2} \log|W| - \frac{1}{2}N(S-P)^T W^{-1}(S-P) \qquad (A5.3).$$

Basically Walker solves the normal equations derived from maximizing the quadratic form in (A5.3) by an iterative method. The major modification employed is to obtain expressions for the elements of W^{-1}. Iterations are continued until convergence is achieved to estimates $\hat{\beta}$ and $\hat{\rho}$ whereupon the M.A. parameters are found from

$$A(L) = \prod_{i=1}^{q} (1-\lambda_i L) \qquad (A5.4)$$

where λ_i are the roots of the equation

$$\sum_{\ell=-q}^{q} \sum_{i,j=0}^{p} \hat{\beta}_i \hat{\beta}_j \hat{\rho}_{\ell-i+j} L^{\ell} = 0 \qquad (A5.5).$$

Therefore the sequence of operations leading to a solution is: solve (A5.3) by any non-linear algorithm to yield estimates $\hat{\rho}, \hat{\beta}$ which are then entered in (A5.5), the roots of the polynomial calculated, and (A5.4) is solved for $\hat{\alpha}$ by equating powers of L.

There are two obvious disadvantages associated with Walker's method. The first is the indirect determination of $\hat{\alpha}$ and the covariance matrix of the estimated parameters. The second is the choice of k. As with Durbin's estimator, it may be shown that the number of covariances required to adequately represent the correlogram of y(t) is a function of the size of the M.A. roots, leading to the conclusion that, although Walker's solution does not require one to fit large A.R.'s, it has a similar deficiency to Durbin's.

(iii) Distributional Properties

Walker presents theorems to the effect that if k is selected so as to include most of the information contained in the correlogram (or

alternatively that $\lim_{N\to\infty} \frac{k}{N} = 0$) then the estimator is consistent, multinormally distributed and efficient. Although a numerical experiment is constructed which affirms this assertion the selected M.A. parameter was $\alpha_1 = 0.5$ and because autocovariances after the sixth yielded little information about the correlation structure of y(t), k could reasonably be set at 5. In summary, Walker's estimator may be a useful tool for ARMA models but the determination of k and the extraction of all estimates is troublesome and militated against its employment in this thesis.

3. Durbin's Frequency Domain Estimator

Durbin [25]has proposed an estimator in the frequency domain which is an analogue to the earlier time domain solution.[1] It derives from the minimization of $\hat{S} = \frac{\hat{I}_{yy}(\lambda)}{f_{yy}(\lambda)}$ where $\hat{I}_{yy}(\lambda)$ is the smoothed spectral density function and $f_{yy}(\lambda)$ has been defined previously. Maximum likelihood estimates of $\theta = (\alpha,\beta)$ may be found from the first order conditions:

$$\hat{I}_{yy}(\lambda)\, f_{uu}^{-1}(\lambda)\, \frac{\partial |B(\lambda)|^2}{\partial \beta_k} = 0 \qquad k = 1,\ldots,p \qquad \text{(A5.6)}$$

$$\hat{I}_{yy}(\lambda)\, f_{yy}^{-2}(\lambda)\, \frac{\partial f_{yy}(\lambda)}{\partial \alpha_k} = 0 \qquad k = 1,\ldots,q \qquad \text{(A5.7)}.$$

For given α (A5.6) is a set of equations linear in β, but for given β (A5.7) is non-linear in α. To overcome this non-linearity Durbin shows that

$$\operatorname{plim}_{N\to\infty} \hat{I}_{yy}(\lambda)\, f_{yy}^{-2}(\lambda)\, \frac{\partial f_{yy}(\lambda)}{\partial \alpha_k} = \hat{I}_{yy}^{-1}(\lambda)\, \frac{\partial f_{yy}(\lambda)}{\partial \alpha_k} \qquad \text{(A5.8)}.$$

---

[1] This account is based on Nicholls [85]. The reader is referred to the more extensive discussion given there.

Replacing (A5.7) by (A5.8) leaves a set of equations with the property that an estimate of either parameter vector may be obtained as the solution of a system of linear equations once the other vector is fixed. This feature suggests the consecutive solution of (A5.6) and (A5.8) (termed the "ping-pong" technique by Aigner [1]) and the sequence of solutions should converge to a limit minimizing $\hat{S}$.[2]

There is little doubt that the method will generalize to ARMAX models and that computationally it may be reduced to a very simple algorithm. However, as Nicholls has emphasized, $\hat{I}_{yy}(\lambda)$ is not unique and, in particular, variation in the number of bands over which the periodogram ordinates are averaged (to obtain the smoothed spectrum) will affect the distributional properties of the estimator. In general there is no way of choosing an optimal bandwidth if the parameters are unknown. One suspects that this difficulty has as its dual the length of the approximating A.R. in the time domain and, if so, the optimal bandwidth would be a function of the roots of the M.A. For this reason it is unlikely that Durbin's estimator will be of use in modelling ARMA/ARMAX processes.

4. Hannan-Nicholls Estimator

Hannan has derived an estimator of ARMA models [44] and Hannan and Nicholls have extended this to ARMAX models [45]. The derivation is based upon the minimization of (2.27) by a Newton-Raphson algorithm

---

[2] Durbin also proposes that (A5.6), (A5.7) be solved directly by inserting consistent estimates of $f_{uu}(\lambda)$ into these equations in place of $f_{uu}(\lambda)$. Nicholls has shown that the variance of the resulting estimates diverges as iteration continues.

modified in that it"..... replaces these second derivatives by corresponding expressions to which they converge almost surely, evaluated at the estimates of the true parameter point. In the case (p) = 0 the method becomes especially simple and these expressions are absolute constants independent of the parameter point.....However if an initial estimate is a long way from the true parameter point the replacement of the second derivatives by these expressions may slow the convergence or might lead to divergence." [45 p.31].

Three factors constitute the case against adoption of the Hannan-Nicholls' estimator.

(i) The derivation of the asymptotic expression for the second derivatives requires a good deal of mathematical skill and as these authors have not considered general models in which non-linearities appear between β and γ this would have to be done in each instance.

(ii) It is not easy to write a program that will incorporate their iterative procedure, and any such program is likely to be inflexible.

(iii) As mentioned above there is some doubt about the robustness of the estimator under poor starting values (see Nicholls [85] for a further discussion of this for a first order moving average model), which coupled with the observation that it is not capable of absorbing advances in numerical methods easily, must mean that convergence will always remain a problem. The incidence of poor starting values is likely to be high in many econometric applications and it is felt that this is a potent reason for seeking modes of attack that are flexible enough to benefit from research designed to provide solutions to this dilemma. It is only fair to point

out that the Hannan-Nicholls' estimator has a number of attractive features, the outstanding one being the provision of central limit theorems guaranteeing M.L. properties for the estimates even if the disturbances are i.i.d. and not n.i.d. Finally some experiments are contained in the joint paper which show that for N=100 the sampling distributions accord well with asymptotic theory, but that for N=40 this was not so. To some extent these results have only a limited significance for economists as the type of filter chosen to generate the exogenous variables results in spectra that are atypical of economic time series.

Appendix 6.

Relationship between the Box-Jenkins and Phillips' Estimators

We begin from Box and Jenkins' Appendix A7.4 [9 p.269-271]. The model set out there is assumed to be generated by a stationary moving average model of order q

$$\tilde{w}(t) = a(t) - \theta_1 a(t-1) - \theta_2 a(t-2) - \ldots - \theta_q a(t-q) \qquad (A6.1).$$

In the terminology of Chapter 2 the following equivalences exist

$$\tilde{w}(t) = u(t)$$

$$a(t) = e(t)$$

$$\theta_k = \alpha_k \qquad k = 1,\ldots,q.$$

Box and Jenkins define the (N+q) dimensional vector $a^T = (a_{1-q}, a_{2-q}, \ldots, \varepsilon_N)$ and the q dimensional vector of preliminary values $a_*^T = (a_{1-q}, a_{2-q}, \ldots, a_0)$ which correspond to the vectors of Chapter 2 as follows

$$a = \begin{bmatrix} e^* \\ e \end{bmatrix} \qquad a_* = e^*.$$

They then define the relationship of (A6.2)(where the definitions of L and X may be found in their Appendix),

$$a = L w_n + X a_* \qquad (A6.2)$$

and obtain the M.L. estimates $\hat{a}_*$ from a minimization of $S(\theta, a_*)$ where $S(\theta, a_*)$ is their equation (A7.4.3)

i.e. $S(\theta, a_*) = (L w_n + X a_*)^T (L w_n + X a_*) \qquad (A6.3).$

Using (A6.2) we see that $S(\theta, a_*) = a^T a$ and from the equivalence above

$$S(\theta, a_*) = e^T e + e^{*T} e^* \qquad (A6.4)$$

which establishes the assertion that this is the sum of squares function

that Box and Jenkins minimize.

The first order conditions for a minimum to S with respect to e* are

$$\frac{\partial S}{\partial e^*} = 2 \frac{\partial e^T}{\partial e^*} e + 2 \frac{\partial e^*}{\partial e^*} e^* \qquad (A6.5).$$

By using the relations $\frac{\partial e}{\partial e^*} = - M^{-1} M^*$, $\frac{\partial e^*}{\partial e^*} = I_q$ (A6.5) reduces to

$$\frac{\partial S}{\partial e^*} = - 2 (M^{-1}M^*)^T e - 2 I_q e^* \qquad (A6.6),$$

whence by equating (A6.6) to zero and substituting $e = M^{-1} (u - M^*e^*)$ we obtain

$$\hat{e}^* = (M^{*T} V^{-1} M^* + I_q) M^{*T} V^{-1} u \qquad (A6.7)$$

as in the text.

The difference between the Box Jenkins and Phillips estimate e* is most easily seen in a first order M.A. case. Consider the ARMA (p,1) model which may be written as

$$u(t) = (1 + \alpha L) e(t).$$

From the concentration formulae we must compute

$$\hat{e}_P = (M^{*T} V^{-1} M^*)^{-1} M^{*T} V^{-1} u \qquad (A6.8),$$

$$\hat{e}_{BJ} = (M^{*T} V^{-1} M^* + I_q)^{-1} M^{*T} V^{-1} u$$

or defining $Z = M^{-1} M^*$, $w = M^{-1} u$

$$\hat{e}_P = (Z^T Z)^{-1} Z^T w \qquad (A6.9).$$

$$\hat{e}_{BJ} = (Z^T Z + I_q)^{-1} Z^T w$$

Now for a first order M.A. (A6.9) has the simple form

$$\hat{e}_P(-1) = \frac{\sum_{j=0}^{N} z(j)w(j)}{\sum_{j=0}^{N} z^2(j)}$$

$$\hat{e}_{BJ}(-1) = \frac{\sum_{j=0}^{N} z(j)w(j)}{1 + \sum_{j=0}^{N} z^2(j)} \qquad \text{(A6.10).}$$

To simplify the derivation we assume that $N \to \infty$. Now from the definition of M* given in Chapter 2 $M^{*T} = [\alpha, 0, \ldots]$ and exploiting the structure of M we obtain a difference equation for z(t)

$$z(t) + \alpha\, z(t-1) = 0 \qquad t = 1, \ldots, \infty$$

$$z(0) = \alpha \qquad \text{(A6.11).}$$

The solution to (A6.11) is

$$z^T = [\alpha, -\alpha^2, \alpha^3, -\alpha^4 \ldots\ldots\ldots]$$

so that $\sum_{j=0}^{\infty} z^2(j) = \alpha^2 + \alpha^4 + \ldots\ldots$

or $$\sum_{j=0}^{\infty} z^2(j) = \frac{\alpha^2}{1 - \alpha^2} \qquad \text{(A6.12).}$$

Similarly $$1 + \sum_{j=0}^{\infty} z^2(j) = \frac{1}{1 - \alpha^2} \qquad \text{(A6.13).}$$

Equations (A6.12) and (A6.13) provide the denominators of the two values of $\hat{e}(-1)$ in (A6.10).

Turning to the computation of $Z^T w$ we commence from the difference equation

$$w(t) + \alpha_1\, w(t-1) = u(t) \qquad t = 1, \ldots, \infty$$

$$w(0) = u(0)$$

with solutions

$$\begin{aligned} w(0) &= u(0) \\ w(1) &= -\alpha\, w(0) + u(1) \\ &\vdots \\ w(k) &= -\alpha\, w(k-1) + u(k) \\ &\vdots \end{aligned} \qquad \text{(A6.14).}$$

By successive substitutions we may reduce (A6.14) to a set of equations in u(t) alone.

$$\begin{aligned} w(0) &= u(0) \\ w(1) &= -\alpha\, u(0) + u(1) \\ w(2) &= \alpha^2 u(0) - \alpha\, u(1) + u(2) \\ w(3) &= -\alpha^3 u(0) + \alpha^2 u(1) - \alpha\, u(2) + u(3) \\ &\vdots \\ w(k) &= -\alpha^k u(0) + \alpha^{k-1} u(1) - \alpha^{k-2} u(2) + \ldots + u(k) \text{ if } k \text{ is odd.} \\ &\vdots \end{aligned}$$

Then it is easily shown that $\sum_{j=0}^{\infty} z(j)\, w(j)$ is

$$\begin{aligned} \sum_{j=0}^{\infty} z(j)w(j) &= u(0)(\alpha + \alpha^3 + \ldots) - u(1)(\alpha^2 + \alpha^4 + \ldots) \\ &\quad + u(2)(\alpha^3 + \alpha^5 + \ldots) - u(3)(\alpha^4 + \alpha^6 + \ldots). \end{aligned}$$

Taking α from the first bracket, α^2 from the second, α^3 from the third and so on yields

$$\begin{aligned} \sum_{j=0}^{\infty} z(j)w(j) &= \alpha\, u(0)(1+\alpha^2+\ldots) - \alpha^2\, u(1)(1+\alpha^2+\ldots) + \alpha^3 u(2)(1+\alpha^2+\ldots) \\ &\quad - \ldots\ldots\ldots\ldots \\ &= \frac{\alpha}{1-\alpha}\left[u(0) - \alpha\, u(1) + \alpha^2\, u(2) - \alpha^3\, u(3) \ldots\ldots\right] \\ &= \frac{\alpha}{1-\alpha^2}\, u^*(0) \end{aligned} \qquad \text{(A6.15).}$$

Now by combining (A6.12), (A6.13) and (A6.15) we obtain

$$\hat{e}_P(-1) = \frac{u^*(0)}{\alpha} \qquad \text{(A6.16)}$$

$$\hat{e}_{BJ}(-1) = \alpha\, u^*(0) \qquad \text{(A6.17)}$$

as is given in the text.

Readers may verify that an application of (A6.17) to the numerical example given on p.214 of [9] with $\alpha = -0.5$ will yield $\hat{e}_{BJ}(-1) = 1.6$ thereby indicating the coincidence of concentration and their forecasting method which they prove must yield $\hat{e}_{BJ}(-1)$.

CHAPTER 3. Monte Carlo Studies of the Phillips Estimator

3.1 Introduction

Most research into the properties of ARMA/ARMAX model estimators has been concerned with the derivation of the moments of the asymptotic distributions. As surveyed in the previous chapter important contributions to this area have been those of Walker [114] and Whittle [116] who considered the properties of estimators that maximised the ARMA likelihood and Pierce [91] and Hannan-Nicholls [45] who concentrated upon the Box-Jenkins and Hannan-Nicholls estimators respectively.

The conclusion to be drawn from the above articles is that the familiar properties of the M.L. estimator of models with non-correlated residuals extend to models in which the residuals are time dependent (see Aitchison and Silvey [2] for a statement of these). However this body of theory is predicated on large samples of data and as econometricians are typically forced to extract information from short time series, asymptotic theorems may not necessarily be relied upon as a guide to suitable estimation procedures. Therefore it is of importance to discover the modification of large sample distributional properties in the translation to small samples and it is this topic that the current chapter is involved with.

Recently there has been a concerted attack upon the derivation of exact finite sampling distributions for a number of estimators. Unfortunately while the models examined have been simple the mathematical skills employed have not and one is led to believe that the Monte Carlo

(or distributional sampling) technique provides the easiest method of assessing finite sampling properties. Owing to the extensive use of this technique in the thesis it is worthwhile noting some of the benefits to be served by conducting such studies.

(1) Information is yielded concerning the small sample distribution of estimators of a model with a given set of parameters.

(2) Insight is provided for the experimenter into the complexities that may be encountered in translating theoretical constructs into practical tools. These insights may be into either the computational sphere or the behaviour of an estimator when the assumption underlying its theoretical derivation is broken e.g. violation of the root restrictions of Chapter One.

(3) Finally it is possible to consider the influence of a wider range of alternative assumptions about the generation of series. An outstanding example of this is non-stationarity. It may be very difficult to establish mathematically the sampling properties of an estimator unless severe restrictions are placed upon the evolutionary nature of the series being analysed, but a Monte Carlo investigation can easily duplicate such a series and extract the distribution. Obviously knowledge that this property holds for a parameter set is very important and certainly better than having no theoretical proof for any parameter set. Two examples in this chapter are non-stationarity in the exogenous variable and the distribution of ARMAX model estimators (which as we have noted in the previous chapter have still not been derived asymptotically).

It will become apparent in this and subsequent chapters that many more Monte Carlo studies were constructed than could be justified

on the basis of (1) alone but it is felt that the primitive state of knowledge on the role of non-linear estimators in time series models made this necessary. Certainly empirical estimation could never have been contemplated without the insights that came from the analysis of simulated data.

3.2 The Design of the Sampling Experiments[1]

An account must be given of the methodology of data generation before any examination of the Monte Carlo experiments. Four criteria guided the construction of synthetic time series. These embody important features of applied econometric models and (hopefully) adherence to these principles will mean that synthetic series resemble actual series.

A. Spectral Characteristics: There is now a considerable body of literature in existence relating the likely spectral shapes of economic variables observed at varying intervals, and the synthetic series should have spectra resembling those that appear in this literature. In particular we will require a concentration of power at the origin to represent annual and/or deseasonalized data and peaks at the various harmonics of a seasonal frequency for series subject to seasonal variation.

B. Lag Distributions: Much econometric research in recent years has centred upon the estimation of distributed lag models and it has become apparent that smoothly declining lag distributions (à la Koyck) are too rigid a specification for some economic relationships e.g. the accelerator, and that this tendency is strengthened the shorter the period of

---

1 The principles stated in this chapter are adhered to throughout the remainder of the thesis and are pertinent to the following chapter and parts of Chapter 6.

observation. For this reason it is necessary to select transfer functions that reflect these studies and this will be done either by a selection from the literature or by ensuring that the theoretical lag distributions have a single mode away from the origin.

C. Correlation between Series: Most econometric models are characterized by an R^2 in the range 0.85-0.95. Although the properties of the statistic are different in a time series situation than in a classical regression it serves as a useful guide to the "unreasonableness" of the assumed relation.

D. Polynomial Roots: Experience indicates that the roots of $B(L)=0$ are likely to have modulus close to unity in many instances so that experiments should be conducted with such a restriction in mind. There is little evidence available concerning the probable magnitude of the roots of $A(L) = 0$ and that which does exist is contradictory e.g. Trivedi's study [107] yields roots close to unity and Williams' [118] first order M.A. parameter never exceeds 0.4. Again resort will be had to the literature for solutions to $B(L) = 0$ and we will generally select roots around 2 for $A(L) = 0$.

Taking the model to be estimated as

$$B(L)y(t) = C(L)x(t) + A(L)e(t)$$

(where $C(L) = 0$ reduces this to an ARMA model) the previous tabulation may be continued with details concerning the generation of data.

E. Properties of the Error Term: The variance of e(t) was set at 3 for all ARMA models and I.B.M. supplied routines (GAUSS and RANDU) were used to generate n.i.d. and i.i.d. variables respectively. A description of these routines may be found in [54].

F. Starting Values: All series were computed recursively by initializing them at zero. This meant that a constant term is not included in any model and the tabulated biases will be smaller than would be the case if a constant term was estimated from the data.

G. Generation of x(t): Two schemes were adopted for the construction of x(t).

A. $(1-\delta_1 L)(1-\delta_2 L^k)x(t) = (1+\delta_3 L)(1+\delta_4 L^m)\varepsilon(t)$

B. $(1-\delta_1 L)x(t) = \varepsilon(t)$.

By varying $k, m, \delta_1, \delta_2, \delta_3$ and δ_4 a considerable change in the nature of x(t) could be induced. We will identify the four parameter combinations that were used most frequently with the understanding that specific mention will be made whenever a filter differing from these is used.

Scheme A.

A.(i) $\delta_1 = 0.95$, $\delta_2 = 0.75$, $k = 1$; $\delta_3 = 0.8$, $\delta_4 = 0.0$, $m = 0$.

A.(ii) $\delta_1 = 1.0$, $\delta_2 = 1.0$, $k = 4$; $\delta_3 = -0.082$, $\delta_4 = -0.372$, $m = 4$.

Scheme B.

B.(i) $\delta_1 = 0.6$.

B.(ii) $\delta_1 = 0.9$.

A(i) was suggested by Grether and Nerlove [36] as an appropriate description of a time series containing only a trend-cycle component and A(ii) is a parametric representation of the quarterly series Household Appliances and Equipment (see Chapter 5). Both filters produce spectral shapes with a strong concentration of power at the origin and additionally A(ii) has minor peaks at $\pi/2$ and π. Such spectra are typical of annual or deseasonalized data (A(i)) and non-deseasonalized

quarterly series (A(ii)) so that the variation in spectral shape conforms to that expected in the data manipulated by economists.

As well as the shape however it is imperative to inspect the level of the spectra. A rough guide may be had from a calculation of the ratios of the spectrum of x(t) at the zero frequency to the ordinates at $\pi/2$ and π. Table 3.1 presents the order of magnitude of these ratios.

Table 3.1

Ratio of $f_{xx}(0)$ to $f_{xx}(\pi/2)$ and $f_{xx}(0)$ to $f_{xx}(\pi)$ for the various schemes for generating x(t)*

Scheme	$\frac{f_{xx}(0)}{f_{xx}(\pi/2)}$	$\frac{f_{xx}(0)}{f_{xx}(\pi)}$
A(i)	$10^4/1$	$10^6/1$
A(ii)	$10^4/1$	$> 10^6/1$
B(i)	10/1	10/1
B(ii)	$10^2/1$	$10^2/1$

*In the case of A(ii) the ratio involves zero terms in the denominator so that the first frequencies before $\pi/2$ and π were used instead.

Most empirical spectra of untransformed economic data are likely to be distinguished by ratios in the range $10^4/1$ to $10^6/1$ indicating that the smaller ratios would be only characteristic of series that had undergone transformations such as differencing. As such transformations are not uncommon in the econometric literature (e.g. the wages equation of the Wharton School model involves first and fourth differencing; see Howrey [53]) it is important to adopt a wide variety of possible schemes

for x(t). However the above note concerning the relevance of each scheme should be borne in mind when assessing the applicability of later conclusions.[2]

As there are now two series of random numbers (e(t) and ε(t)) a decision must be made on their relative variance. The quantity employed in selecting this ratio (defined as $\sigma_e^2/\sigma_\varepsilon^2$) was the mean value of the correlation coefficient (R^2) over 50 replications. As mentioned previously most time series models possess an R^2 in the range 0.85-0.95 and the variance ratios needed to duplicate this are:-

A(i) : 500/1

A(ii) : 500/1

B(i) : 4/1

B(ii) : 10/1.

These are the values normally adopted and R^2 ranged from 0.9 to as high as 0.99 with a concentration around 0.96.[3]

H. Sample Size: Two factors were influential in the selection of the sample sizes to be examined.

---

[2] At this stage one might as well lay to rest a myth that is prevalent in the literature pertaining to the simulation of time series-namely that filters of type B(i) reproduce Granger's "typical spectral shape" and therefore are realistic. As shown in the text, although the shape of the spectrum is correct, the level is not so that the use of such a filter restricts the generality of any results presented. For examples of its use in this field see Hendry and Trivedi [49], Hannan and Nicholls [45].

[3] To avoid the possibility of a pseudo-infinite variance for σ_e^2, σ_ε^2 was set at 0.01 so that a variance ratio of 500/1 could be achieved by letting $\sigma_e^2 = 5$.

(i) The asymptotic behaviour of an estimator. This concern is occasioned by the paucity of asymptotic theory for estimators of ARMAX models, but as well an awareness of the sample size needed for reliance upon existing asymptotic theory will be created. Such insight is important as most statistics used for hypothesis testing when a lagged dependent variable is present are powerful only in very large samples. To this end samples of one hundred observations were constructed. It is worthwhile mentioning that for some economic time series such sample sizes are the rule rather than the exception e.g. Monthly Bank Advances.

(ii) The small sample behaviour of the estimator. To investigate this two sizes were chosen - 40 and 70 - the justification being that most quarterly (or monthly) time series are likely to be at least of this length and it is in the modelling of these series that the ARMA form is likely to be most effectively employed. The only other known Monte Carlo study of the Phillips estimator - that of Hendry and Trivedi (H-T) [49] - uses samples of size 25, 40 and 50 thereby overlapping our own work.[4] A critique of this study is contained in Appendix 7.

I. Equation Type: Both ARMA and ARMAX equations are simulated in the following sections. Two arguments may be advanced to justify the dual approach.

(i) A number of authors have contended that ARMA models are good forecasting tools and the present writer believes that they are the natural "naive" model against which various econometric specifications may be

---

[4] Actually very few of the results presented in the H-T report duplicate my own because of my interest in the difficulties raised in the application of such models which led to a different perspective to that found in H-T. Where there is a clash H-T's work is given precedence.

tested. This latter theme is developed in the concluding chapter and will not be pursued in any more detail here. Nevertheless if the two utilities mentioned above are ascribed to ARMA models it is of considerable interest to assess the possibility of accurate estimation of the parameters.

(ii) Embodying as they do the twin features of an autocorrelated disturbance term and lagged endogenous variables, ARMA models exhibit those characteristics that invalidate traditional estimators such as O.L.S., and therefore may be regarded as simplified econometric models. A further defence - emphasizing the polar nature of this form - lies in an analogy with an example quoted by Malinvaud [71] in which the absolute magnitude of the inconsistency of the O.L.S. estimator of β_1 in a first order A.R. with autocorrelated disturbances is smaller the greater the (positive) autocorrelation present in the exogenous variable. Note that the inconsistency is merely reduced by the presence of an exogenous variable but never eliminated. Therefore the ARMA model is a polar one. Having established this it is obligatory to recognize that ARMAX models are of greatest importance to an econometrician so that whilst a polar case yields valuable information it must be supplemented by a thorough investigation of the transition to more realistic forms.

J. Data Summary: The processing of all numerical results will be performed with the aid of seven statistics - a set of six which summarize the data and a seventh that acts as a guide to the accuracy of the former set.

(a) The mean over m replications i.e. $\bar{\theta} = m^{-1} \sum_{1}^{m} \hat{\theta}(j)$ where $\hat{\theta}(j)$ is the

value of $\hat{\theta}$ obtained from the j'th replication.

(b) The standard deviation over m replications i.e.

$$\sigma_{\hat{\theta}} = [(m-1)^{-1} \sum_{1}^{m} (\hat{\theta}(j)-\bar{\theta})^2]^{\frac{1}{2}}.$$

(c) The bias $(\bar{\theta} - \theta)$ where θ is the population parameter.

(d) The ratio of the empirical to asymptotic variance i.e. $\sigma^2_{\hat{\theta}}/\sigma^2_{\theta}$ where σ^2_{θ} is computed from the asymptotic formulae of Chapter 2 or Appendix 4. This ratio is distributed as $F_{m-1,\infty}$ if $\sigma^2_{\hat{\theta}} > \sigma^2_{\theta}$, or if $\sigma^2_{\hat{\theta}} < \sigma^2_{\theta}$ the inverse ratio is distributed as $F_{\infty,m-1}$.

(e) Where comparisons between two estimators θ_1 and θ_2 are important the difference between the mean values $\bar{\theta}_1$ and $\bar{\theta}_2$ is tested by forming the ratio $\frac{7(\bar{\theta}_1 - \bar{\theta}_2)}{(\sigma^2_{\hat{\theta}_1} + \sigma^2_{\hat{\theta}_2})^{\frac{1}{2}}}$ and testing this as a t statistic with 98 degrees of freedom. The statistic is formed from Mood and Graybill [75 p.306] by setting m = n = 50 and its validity hinges on the normality of $\hat{\theta}_1$ and $\hat{\theta}_2$ - a feature that is suspect in the smaller samples but which might be expected to hold in samples of 100. Additionally it must be assumed that there is a common population variance.

(f) Under the same conditions as (e) a test on the equality of the variances is available by forming the ratio σ^2_1/σ^2_2 and regarding it as an $F_{49,49}$ variable.

(g) Because the Monte Carlo technique is itself subject to statistical fluctuation (i.e. the statistics calculated from (a) - (f) above are from empirical distributions which converge to the true sampling

distributions only as the number of replications tends to infinity) statistical tests should be performed on these measures to assess their reliability. The principal device exploited for this purpose was the standard error of the bias $\sigma_{(\bar{\theta}-\theta)} = m^{-\frac{1}{2}} \sigma_{\hat{\theta}}$ and the statistic $(\bar{\theta}-\theta)/\sigma_{(\bar{\theta}-\theta)}$ is distributed as Student's t with (m-1) degrees of freedom.

Concern over the size of m is a feature of many Monte Carlo studies and (g) illustrates why this is so. Clearly the sampling error of the statistics will vary inversely with m and at least for the bias the inverse relationship is with the square root of m signifying that a large number of replications is mandatory before there can be satisfaction concerning the accuracy of the measures computed from (a) - (f). As a policy of increasing m may be very expensive in terms of computer time a number of techniques such as antithetic variables (see [39] for a discussion) have evolved, which are designed to "effectively" increase m without involving a proportionate amount of computing. Ideally such methods should be applied in all Monte Carlo work but it is felt that setting m = 50 was sufficient to yield reliable conclusions. In those instances when m could not be raised above 20 the resulting conclusions must be viewed with some scepticism.

3.3. An Overview of the Remaining Sections

Here we present a grouping of the remaining sections. As the objective of this chapter is to isolate sampling properties under varying conditions, it was decided to group the studies into six separate sections encompassing some of the problems arising in the application of these estimators. The six are:-

1. Spectral Shape: It is proposed to examine the sensitivity of the estimator to a wide variety of spectral shapes for the A.R. and M.A. in order to gain an appreciation of the degree of success in the application of an estimator to any economic problem.

2. Misspecification and Lack of Identification: Misspecification is a universal problem in econometric modelling but has added perplexities in ARMAX models because of a possible lack of identification. The sensitivity of the estimator to misspecification and the forces leading to a lack of identification will be examined.

3. Rational Lags and Expectations: Chapter 1 has noted that the genesis of ARMAX models may be the assumption of rational spectral density functions. The hallmark of such models is the restrictions that exist between the A.R. and M.A. parameters and it is profitable to enquire into the statistical properties of estimators that exploit or fail to exploit these restrictions.

4. Seasonal Models: As the sample sizes were selected to reflect the availability of quarterly observations it is important to conduct some experiments which reflect the special characteristics of this data. In particular we are concerned with the seasonal patterns observed in most economic time series collected at this time interval. Such "periodic" behaviour impinges upon our models in either of two ways. Firstly the exogenous variable may reflect a seasonal influence. Secondly even if this is not so (or all data has been seasonally adjusted) the disturbance term, by absorbing the influence of omitted variables, may have peaks in its spectrum at the seasonal frequencies.

Both possibilities will be studied in this section.

5. Departures from Stationarity: As no vestiges of a proof are available concerning the asymptotic properties of ARMAX estimators when the assumption of covariance stationarity of x(t) is broken it is elemental that some investigation should be made of this, especially in view of the suspicion that many economic time series fall into this category.

6. Departures from Normality: A central feature of the maximum likelihood estimator in the classical regression situation is that it is consistent and asymptotically normally distributed even when the disturbances are i.i.d. rather than n.i.d. thereby widening the class of disturbance processes that may be allowed. Hannan and Nicholls have provided a Central Limit Theorem showing that their estimator has this property and Walker has sketched a similar proof for ARMA model estimators. Faced with these general results it was decided to test the sensitivity of some of the conclusions reached in the previous sections by using white rather than Gaussian noise in the simulations.

3.4 Effects of Spectral Shape

(1) ARMA Models

Five ARMA models, representing a wide range in the spectral shape of each component process, and thereby the overall spectral density, were selected. These are labelled Models (A) - (E) and are as follows.

MODEL A

$$(1-0.8L)y(t) = (1+0.5L)\, e(t)$$

MODEL B

$$(1-1.3L + 0.6L^2)\, y(t) = (1+0.6L + 0.3L^2)\, e(t)$$

MODEL C

$$(1-1.57L + 1.23L^2 - 0.42L^3)\, y(t) = (1+1.2L + 0.62L^2 + 0.27L^3)\, e(t)$$

MODEL D

$$(1-1.031L + 0.630L^2 - 0.524L^3)\, y(t) = (1+0.8L + 0.3L^2)\, e(t)$$

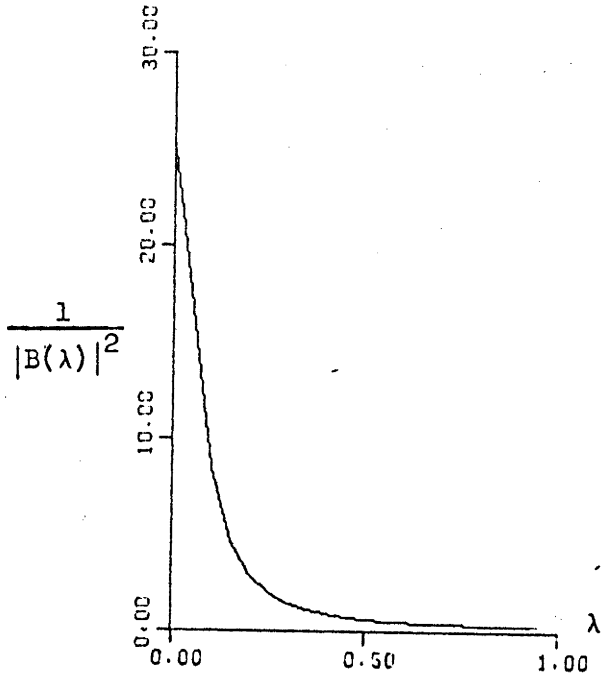
MODEL E

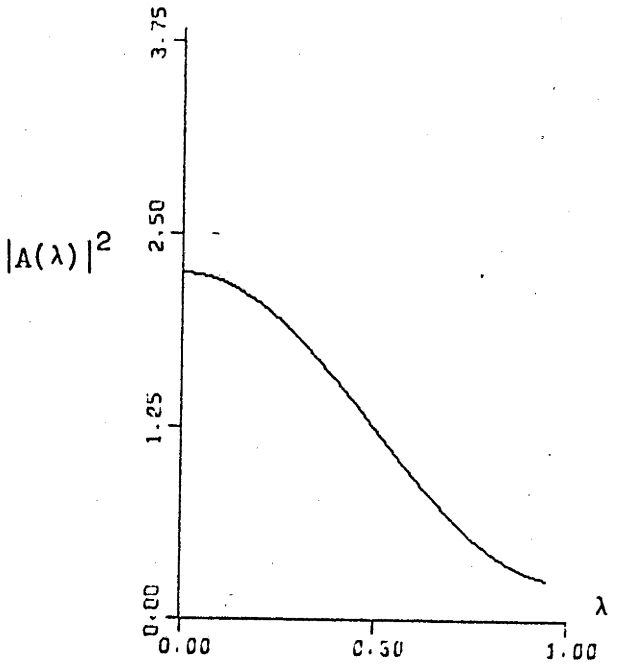
$$(1-1.55L + 0.6L^2)\, y(t) = (1+0.8L + 0.3L^2)\, y(t)$$

Appendix 8 provides a listing of the roots of $B(L) = 0$, $A(L) = 0$ of Models (A) - (E) (and also for all other transfer functions used in this chapter) and from this it is clear that Models D and E possess the most likely roots for untransformed economic series. Figures 3.1 - 3.5 present a plot of the squared gains of the frequency response functions and the spectral density function of $y(t)$ associated with each of the five models.

FIGURE 3.1

SPECTRAL CHARACTERISTICS OF MODEL A





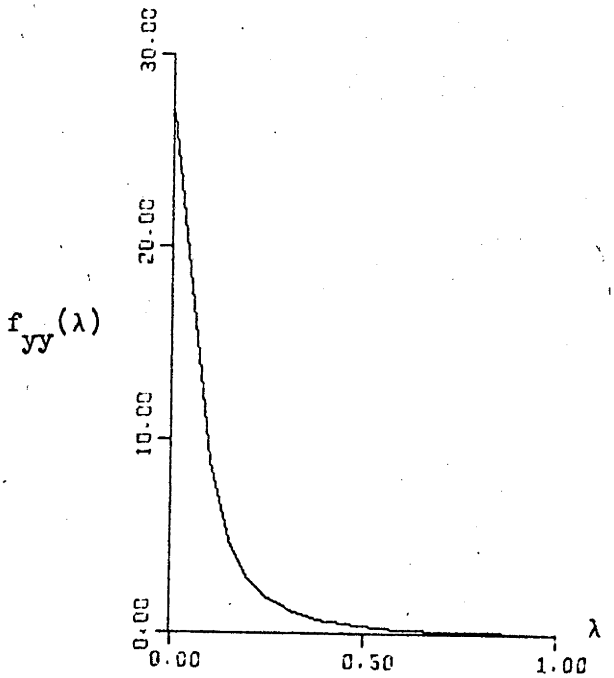
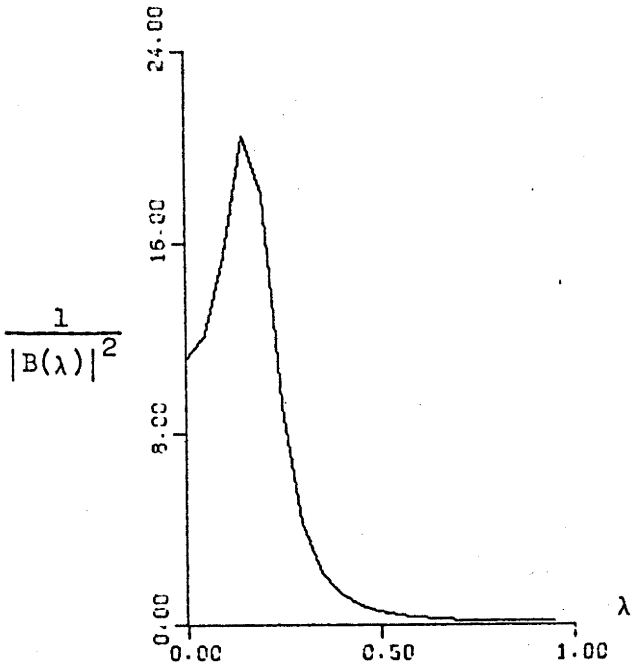
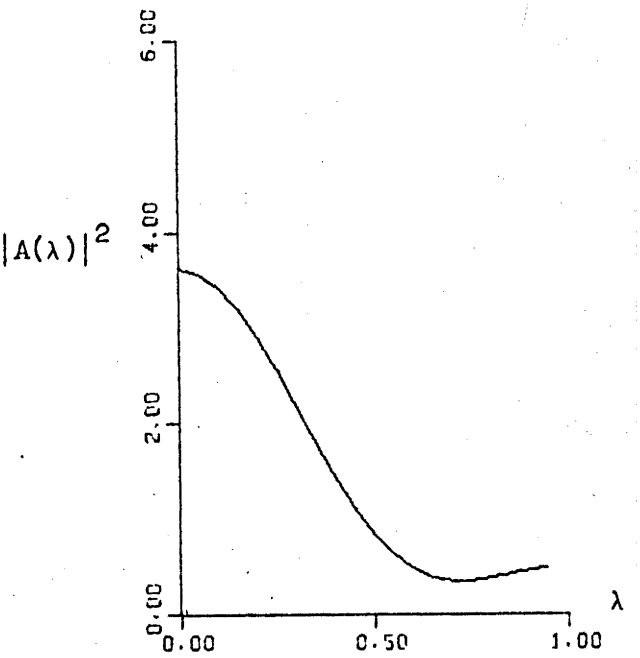


FIGURE 3.2

SPECTRAL CHARACTERISTICS OF MODEL B





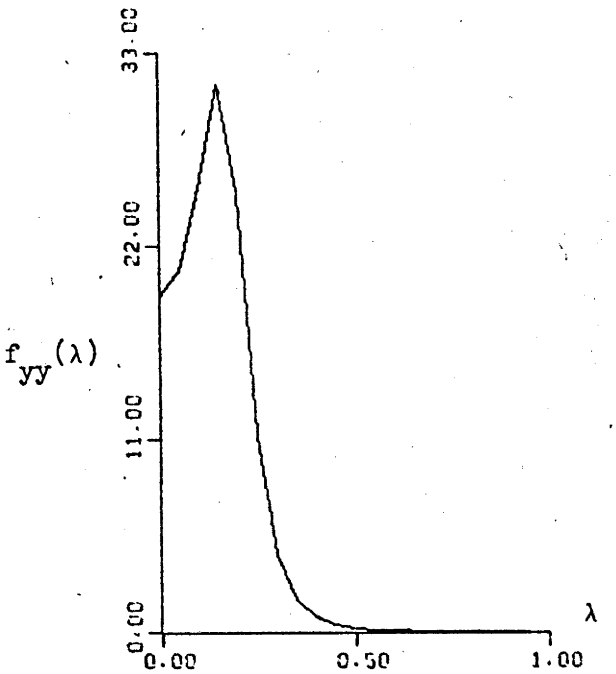
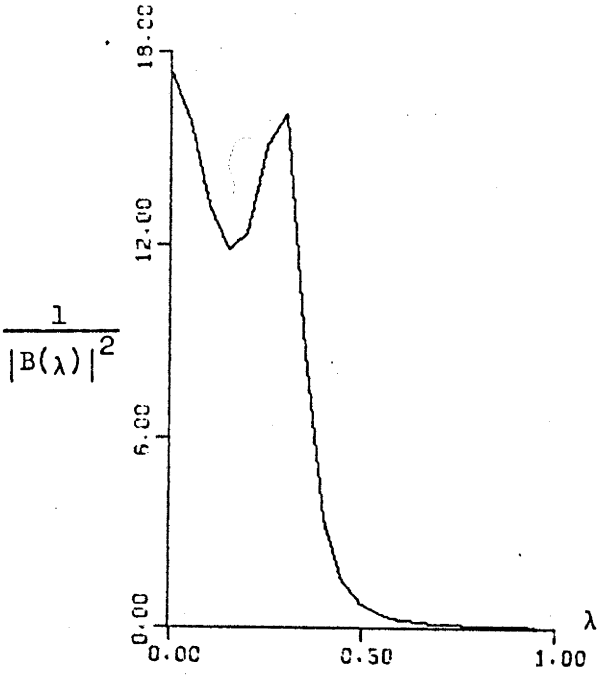
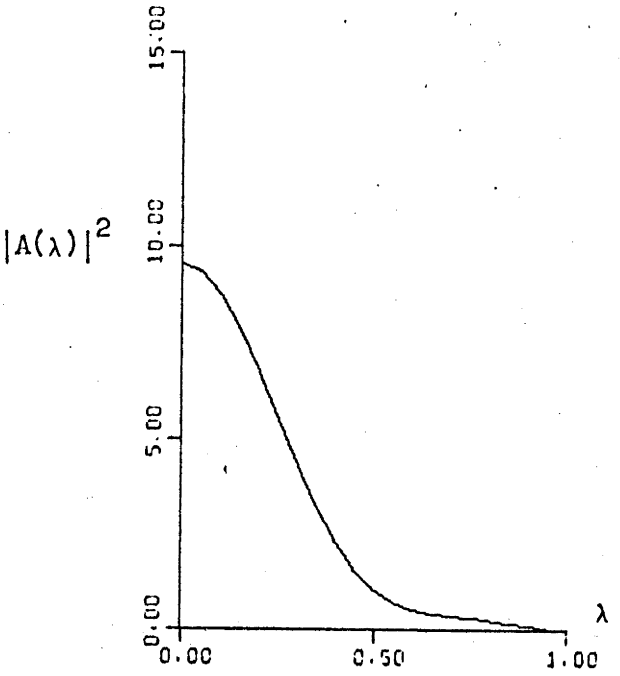


FIGURE 3.3

SPECTRAL CHARACTERISTICS OF MODEL C





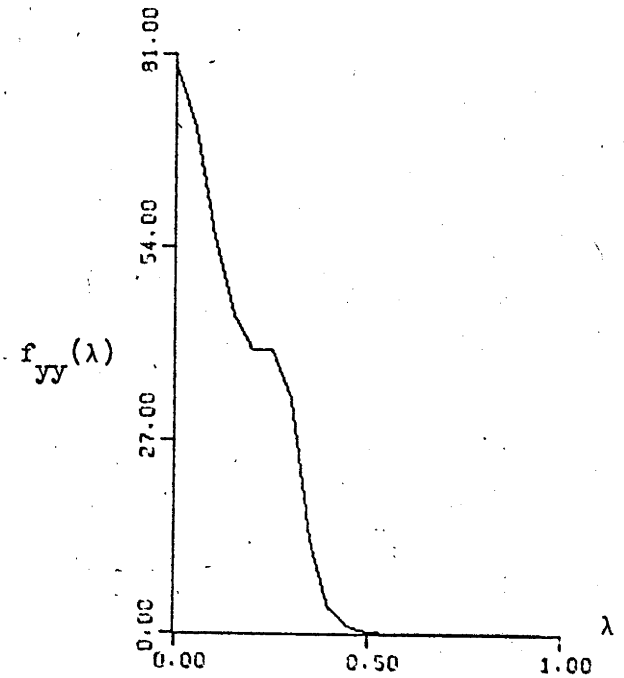
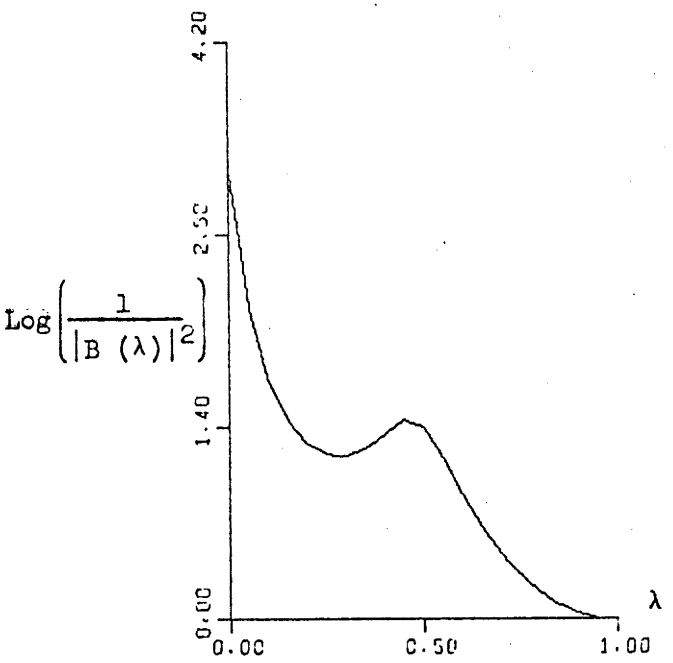
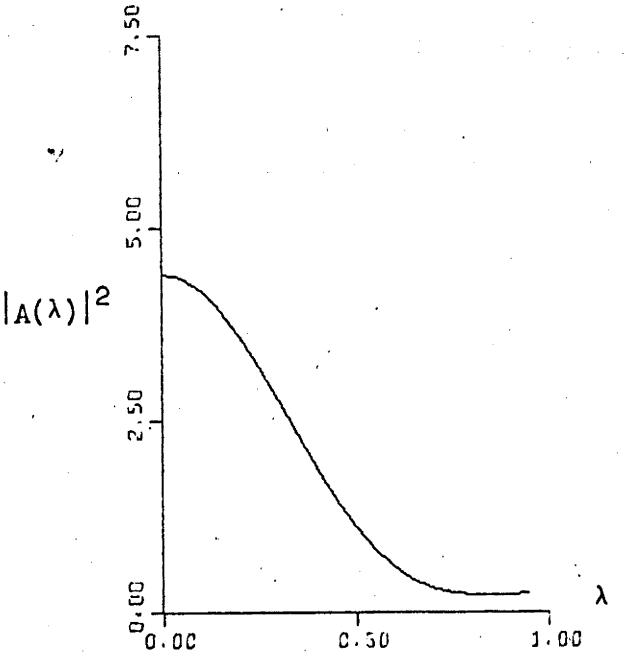


FIGURE 3.4

SPECTRAL CHARACTERISTICS OF MODEL D





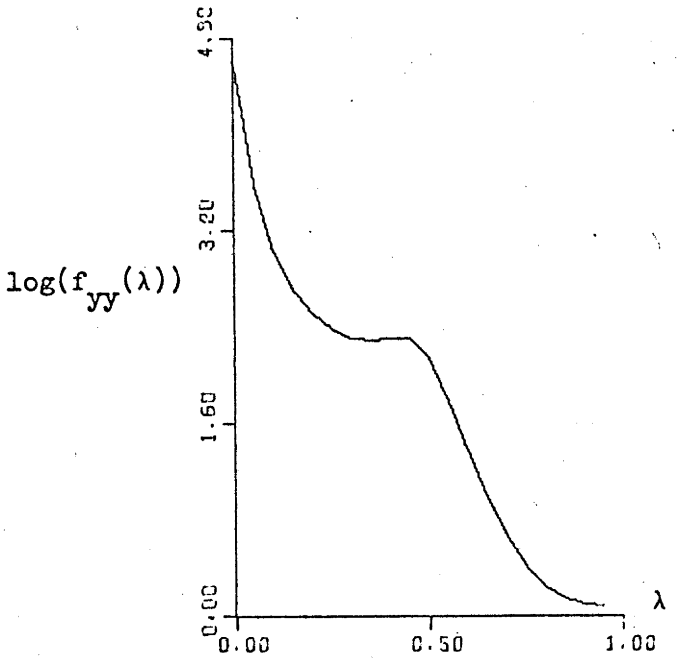
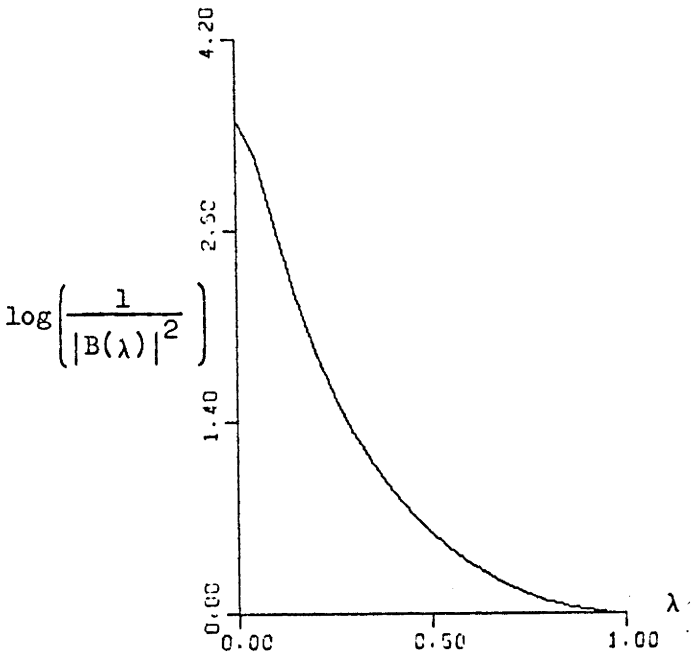
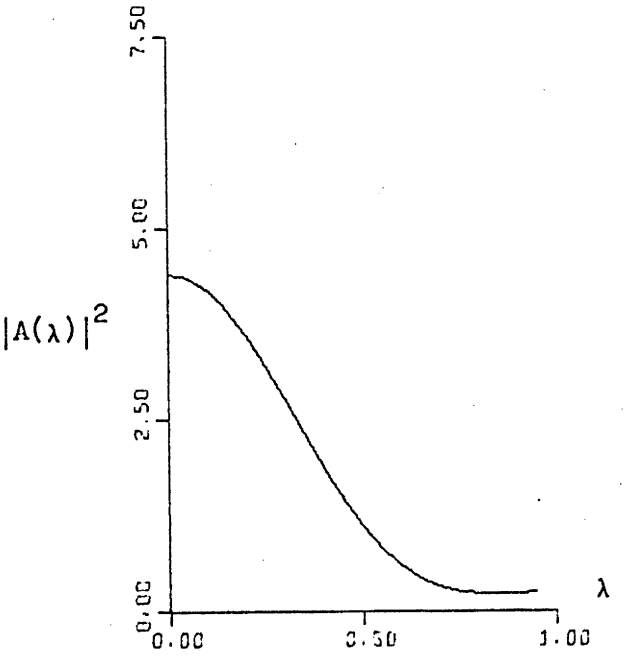
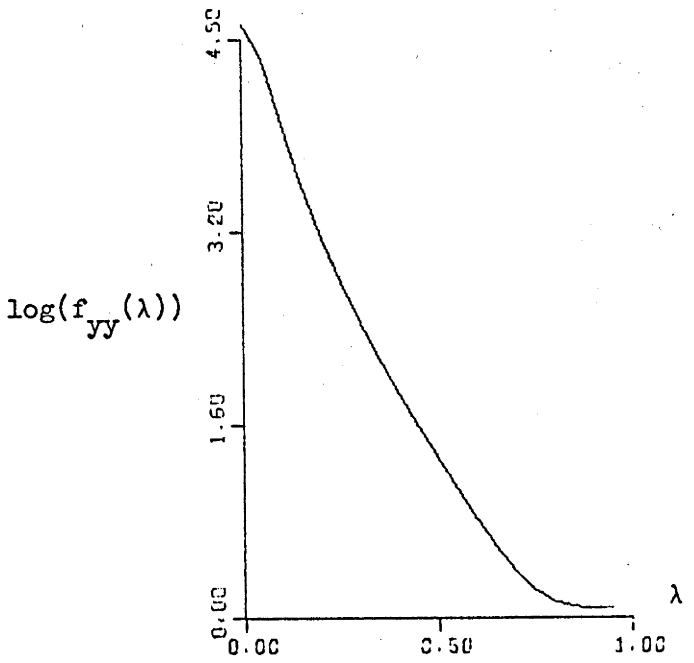


FIGURE 3.5

SPECTRAL CHARACTERISTICS OF MODEL E







The plotted functions provide a convenient summary of the salient features of each model and suggest that Models A and E may be regarded as representative of unfiltered yearly data or deseasonalized quarterly data while Model D with its peak at $\pi/2$ in the squared gain of the A.R. response function will yield a series with spectral density similar to a non-deseasonalized quarterly time series. The remaining two models are characterized by squared gain functions with peaks at $\pi/5$ (Model B) and $\pi/3$ (Model C) respectively. In the case of Model C this is not transmitted to $f_{yy}(\lambda)$ but $y(t)$ of Model B exhibits cycles around $\pi/5$. Is it possible to observe actual economic time series that behave in this fashion? The answer to this is dependent upon the extent of prior filtering of the data and/or the form in which the dependent variables appear (except that Model B is close to the spectrum of the sunspots data; see Anderson [4]).

Examples of prior filtering are contained in:

(a) The Macroeconometric literature e.g. the attenuation of frequencies induced by differencing and the use of ratios and rates of growth as dependent variables.

(b) The time series literature e.g. Nerlove's spectral study of first differenced price data [82] and Howrey's report of the spectra of variables included in the Wharton School Model [53]. There seems no reason to suppose that the M.A. transfer functions would be an unrealistic description of the error process if seasonal variation has been adequately removed and the specification is accurate.

Tables 3.2, 3.3 and 3.4 present the summary statistics for Models (A) - (E) for the three sample sizes N = 40, 70 and 100. The column headings (adhered to throughout the thesis) are

N - sample size

P - True Parameter Values

M - Mean (Statistic (a) of Section 3.2)

B - Bias (Statistic (c))

S.E. - Standard Error of the Bias
B/SE - The absolute value of the ratio of B to S.E. } statistic (g)

S.D. - Standard Deviation (statistic (b))

A.S.D.- Asymptotic Standard Deviation
SD/ASD- ratio of SD to ASD } statistic (d)

Table 3.2

Summary Statistics for Models (A) - (E)

Sample Size N = 40

Model	P	M	B	SE	B/SE	SD	ASD	SD/ASD
A	0.8	0.741	-0.059	0.0180	3.28	0.1264	0.1048	1.21
	0.5	0.550	0.050	0.0241	2.07	0.1692	0.1513	1.12
B	1.3	1.2665	-0.0335	0.0311	1.08	0.2198	0.1901	1.16
	-0.6	-0.5980	0.0020	0.0296	0.07	0.2092	0.1759	1.19
	0.6	0.6256	0.0256	0.0443	0.58	0.3129	0.2183	1.43
	0.3	0.3579	0.0579	0.0428	1.35	0.3026	0.1991	1.52
C	1.57	1.3483	-0.2217	0.0564	3.93	0.3989	0.3494	1.14
	-1.23	-0.9435	0.2865	0.0775	3.70	0.5479	0.4705	1.16
	0.42	0.2523	-0.1677	0.0460	3.65	0.3252	0.2636	1.23
	1.20	1.4709	-0.2709	0.0632	4.29	0.4474	0.3578	1.25
	0.62	0.9746	0.3546	0.0933	3.80	0.6598	0.5118	1.29
	0.27	0.4199	0.1499	0.0488	3.07	0.3452	0.2504	1.38
D	1.031	0.9388	-0.0922	0.0468	1.97	0.3312	0.2816	1.18
	-0.630	-0.6037	0.0263	0.0526	0.50	0.3722	0.3373	1.10
	0.524	0.5217	-0.0023	0.0249	0.09	0.1763	0.1619	1.09
	0.8	0.9046	0.1046	0.0579	1.81	0.4091	0.3137	1.30
	0.3	0.4524	0.1524	0.0473	3.22	0.3344	0.2726	1.23
E	1.55	1.4250	-0.1250	0.0361	3.46	0.2553	0.1772	1.44
	-0.6	-0.4976	0.1024	0.0351	2.92	0.2485	0.1754	1.42
	0.8	0.9320	0.1320	0.0468	2.82	0.3310	0.2057	1.61
	0.3	0.4514	0.1514	0.0455	3.33	0.3218	0.1953	1.65

Table 3.3

Summary Statistics for Models (A) - (E)

Sample Size N = 70

Model	P	M	B	SE	B/SE	SD	ASD	SD/ASD
A	0.8	0.759	-0.041	0.0129	3.18	0.0905	0.0784	1.15
	0.5	0.544	0.044	0.0169	2.60	0.1182	0.1131	1.05
B	1.3	1.2978	-0.0022	0.0224	0.10	0.1582	0.1437	1.10
	-0.6	-0.6078	-0.0078	0.0204	0.38	0.1444	0.1329	1.09
	0.6	0.6147	0.0147	0.0298	0.49	0.2105	0.1650	1.28
	0.3	0.3135	0.0135	0.0273	0.49	0.1931	0.1505	1.28
C	1.57	1.4314	-0.1386	0.0473	2.93	0.3343	0.2642	1.27
	-1.23	-1.0421	0.1879	0.0604	3.11	0.4274	0.3558	1.20
	0.42	0.3061	-0.1139	0.0329	3.46	0.2327	0.1993	1.17
	1.20	1.3839	0.1839	0.0478	3.85	0.3377	0.2707	1.25
	0.62	0.8502	0.2302	0.0692	3.33	0.4891	0.3871	1.26
	0.27	0.3714	0.1014	0.0342	2.96	0.2419	0.1893	1.28
D	1.031	0.9590	-0.0720	0.0335	2.15	0.2371	0.2139	1.11
	-0.630	-0.5610	0.0690	0.0391	1.76	0.2766	0.2549	1.09
	0.524	0.4737	-0.0503	0.0241	2.09	0.1512	0.1231	1.23
	0.8	0.8937	0.0937	0.0437	2.14	0.3090	0.2370	1.30
	0.3	0.3704	0.0704	0.0389	1.81	0.2751	0.2060	1.34
E	1.55	1.5255	-0.0245	0.0199	1.23	0.1410	0.1339	1.05
	-0.6	-0.5900	0.0100	0.0202	0.50	0.1428	0.1325	1.08
	0.8	0.8380	0.0380	0.0257	1.48	0.1816	0.1555	1.17
	0.3	0.3211	0.0211	0.0250	0.84	0.1765	0.1476	1.20

Table 3.4

Summary Statistics for Models (A) - (E)

Sample Size N = 100

Model	P	M	B	SE	B/SE	SD	ASD	SD/ASD
A	0.8	0.762	-0.038	0.0109	3.49	0.0763	0.0653	1.17
	0.5	0.541	0.041	0.0134	3.06	0.0942	0.0942	1.00
B	1.3	1.3268	0.0268	0.0166	1.61	0.1172	0.1202	0.98
	-0.6	-0.6034	-0.0034	0.0147	0.23	0.1037	0.1112	0.93
	0.6	0.5752	-0.0248	0.0229	1.08	0.1620	0.1380	1.17
	0.3	0.2651	-0.0349	0.0220	1.59	0.1558	0.1259	1.24
C	1.57	1.4935	-0.0765	0.0405	1.89	0.2864	0.2211	1.30
	-1.23	-1.1201	0.1099	0.0544	2.02	0.3849	0.2978	1.29
	0.42	0.3443	-0.0757	0.0295	2.57	0.2089	0.1668	1.25
	1.20	1.2927	0.0927	0.0426	2.18	0.3015	0.2265	1.33
	0.62	0.7195	0.0995	0.0592	1.68	0.4187	0.3239	1.29
	0.27	0.3142	0.0442	0.0272	1.63	0.1923	0.1583	1.21
D	1.031	0.9563	-0.0747	0.0266	2.81	0.1880	0.1792	1.05
	-0.630	-0.5541	0.0759	0.0316	2.40	0.2232	0.2133	1.05
	0.524	0.4799	-0.0441	0.0149	2.96	0.1053	0.1030	1.02
	0.8	0.8842	0.0842	0.0336	2.51	0.2376	0.1983	1.20
	0.3	0.3480	0.0480	0.0324	1.48	0.2288	0.1723	1.33
E	1.55	1.5369	-0.0131	0.0162	0.81	0.1144	0.1120	1.02
	-0.6	-0.6007	-0.0007	0.0155	0.05	0.1095	0.1109	0.99
	0.8	0.8178	0.0178	0.0199	0.89	0.1406	0.1301	1.08
	0.3	0.2914	-0.0086	0.0202	0.43	0.1428	0.1235	1.16

A close examination of Tables 3.2, 3.3, and 3.4 reveals some interesting features.

1. Models A and D exhibit significant biases in samples of size 70 and 100 (except for α_2(N = 70, 100) and β_2(N = 70) of Model D). Walker has derived the theoretical bias of β_1 in Model A as 0.0422 (under the assumption of mean corrected data) and comparing this with the empirical bias it seems that the development of formulae for bias corrections may be of value in these models.

2. There is a strange pattern to the statistics of Model D - which is likely to recur - and it should be commented upon now. The biases at sample size 40 are insignificant but become significant as the sample size increases. In part this may be attributed to a skewed sampling distribution in small samples making the mean a poor exclusive measure of central tendency as it lies a considerable distance from the mode and median. As well as this however it will be difficult for the estimator to be normally distributed around the true values of the parameters because of the high roots of B(L) = 0: sampling fluctuation and the enforcement of the root restriction will yield a truncated distribution. To judge this, Table 3.5 presents the number of replications at which the restriction was operative for various models and sample sizes.

Clearly the statistics of Model D will suffer most from the imposition of the root restriction and Models B and E the least. This accords with the tables in that for N > 70 both of the latter models are well estimated. Because of the presence of the root restrictions there will certainly be some induced bias in the smaller samples so that we may hypothesize that the larger samples are a better reflection of the significance of biases etc.

Table 3.5

Frequency of Use of Root Restriction

Models B,D and E: All Sample Sizes

Sample Size	B	D	E
40	4	15	3
70	0	5	0
100	0	2	0

3. Model E which has a similar A.R. response function to Model A and the M.A. response function of Model D displays significant biases only in the small sample of 40, and the improvement in the properties of the estimator in the transition from N = 40 to N = 70 is quite remarkable. Certainly there is no evidence of the biases associated with Model A. To resolve this we note that the average $\bar{R}^2$ (defined as $\bar{R}^2=1-\frac{\Sigma e^2(t)}{\Sigma y^2(t)}$) [5] was 0.77 for Model A and 0.96 for Model E and taking this to be a crude signal/noise ratio the former process incorporates more "noise" than the latter and the extraction of the signal may require a very much greater number of observations. To derive some relationships we use the asymptotic equalities.

$$\sigma_y^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{yy}(\lambda)d\lambda = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sigma^2}{2\pi}\,\frac{|A(\lambda)|^2}{|B(\lambda)|^2}\,d\lambda$$

so that

[5] To be distinguished from $R^2 = 1 - \frac{\Sigma u^2(t)}{\Sigma y^2(t)}$.

$$\bar{R}^2 = 1 - \frac{1}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|A(\lambda)|^2}{|B(\lambda)|^2}\, d\lambda}$$

In early experiments [88] (by switching transfer function i.e. the A.R. became the M.A. and vice versa) it emerged that the Phillips estimator was sensitive to the ratio

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|A(\lambda)|^2}{|B(\lambda)|^2}\, d\lambda.$$

Specifically, the biases were larger the smaller the ratio. Reformulated as an effect upon $\bar{R}^2$ it is clear that the theoretical value of $\bar{R}^2$ will vary with this ratio and in such a way that a rise in the ratio will increase $\bar{R}^2$. Therefore it is probable that the small sample bias is a function of $\bar{R}^2$ i.e. a signal/noise ratio.

3. It is possible to contend on the basis of Tables 3.2 - 3.5 that the spectral properties of each model were a determinant of the performance of the estimator. More exactly a distinction is drawn between response functions that are complex and those that are differentiated e.g. Model C exhibits an A.R. response function that is both complex and sharply differentiated from the M.A. response function while Model B is also differentiated but simpler. Therefore the hypothesis is advanced that there is a trade-off between the two factors. To begin the test we note that Model B is better estimated in all sample sizes than Model E and the primary difference in the models is undifferentiated response functions for the latter. Model C on the other hand possesses complex differentiation and biases only tend to insignificance in very large samples. For the sample size N = 100 three parameters had significant bias and four had a significantly inflated variance but the differences

were not great when compared to the situation ruling for N = 40. Model D provides a difficult case to judge. With the exception of the small peak at $\pi/2$ the response functions are very similar so that it may be closer to Model E than B. Additionally the smallness of the peak might lead to the need for a larger number of observations to achieve an adequate resolution since - heuristically anyway - the proportional relationship between the number of periodogram ordinates and sample size would hint that the greater the number and width of peaks, and the less sharp their delineation, the greater the information loss at any sample size. The moral of the section is that caution should be exercised when filtering series in order to avoid the creation of spectra with marked modes.

4. Model D should be given attention for two reasons. Firstly the A.R. and M.A. transfer functions are used repeatedly throughout the remainder of the chapter as constituents of ARMAX models so that we would expect to see some change in the statistics above in the transition to a different form. Secondly the A.R. transfer function was taken from Sargan's study of wages and prices in the United Kingdom [96] and might therefore be regarded as representative of the type of transfer function encountered in modelling economic time series.

5. All of the above discussion has concentrated upon a comparison of asymptotic and sample moments but the estimator should not only possess the correct mean and variance but should also be asymptotically normally distributed. To test this hypothesis the Kolmogorov - Smirnov D statistic was calculated for Models B and E above with the true values of the parameters and the asymptotic variances as the first two moments. Table 3.6 contains the statistic for these models.

Table 3.6

Kolmogorov-Smirnov D Statistic for Models B and E

Model	Sample	D			
		β_1	β_2	α_1	α_2
Limits					
1%		0.2305	0.2305	0.2305	0.2305
5%		0.1923	0.1923	0.1923	0.1923
10%		0.1725	0.1725	0.1725	0.1725
B	40	0.0932	0.1294	0.1924	0.1555
	70	0.0685	0.0885	0.1142	0.1031
	100	0.1466	0.0802	0.1751	0.1492
E	40	0.2460	0.1964	0.2378	0.2499
	70	0.1161	0.0818	0.1394	0.1005
	100	0.1030	0.0793	0.0915	0.1111

Taking a significance level of 5% the hypothesis of normality is only rejected for Model E in sample size 40. This is in agreement with the summary statistics of Tables 3.2 - 3.5 in that the moments of the estimator of Model B parameters were close to their asymptotic values in all samples whereas for Model E there was considerable divergence in the smaller samples. In general it was found that if the first two moments were not significantly different from their asymptotic values then normality held. However an application of the D statistic to all models is foiled by the root restrictions. Almost certainly these reduce the numbers in the tails of the empirical distributions and as the D statistic depends upon the maximum absolute deviation between the theoretical and empirical distributions the statistic is biassed toward an acceptance of the null hypothesis. Therefore it is expected that there would be a tendency to accept normality in small samples as root restrictions are used most frequently then and this may

be the explanation of the acceptance of normality when N = 40 for Model B.

6. Of the five models the asymptotic variance formulae were applicable for four at sample size 70 and for one at the very small sample of 40. However it would be unwise to extend this conclusion to all models - as Model C shows - and the most general principle to be extracted from this sub-section might be that the application of ARMA estimators to series with less than 70 observations should be discouraged if one wishes to appeal to asymptotic theorems.

(ii) ARMAX Models

Models F and G are selected to illustrate the performance of the Phillips estimator in the context of ARMAX Models.

Model F

$$(1-1.031L + 0.630L^2 - 0.524L^3)\, y(t) = -0.375\, x(t) + (1+0.8L + 0.3L^2)\, e(t)$$

Model G

$$(1-1.295L + 0.428L^2)\, y(t) = 1.0\, x(t) + (1 + 0.5L)\, e(t)$$

The response functions associated with Model F have been discussed before and both of Model G have a smooth monotonically declining shape. The A.R. transfer function of Model G is to be found in Jorgenson's study of investment demand for the Brookings Model [58]. There are peaks in the lag distributions implied by F but not by G (see Appendix 8 for these and the roots).

Tables 3.7 and 3.8 give the summary statistics for each of the models for the three sample sizes.

Table 3.7

Summary Statistics for Model F

All Sample Sizes

N	P	M	B	SE	B/SE	SD	ASD	SD/ASD
40	1.031	1.0082	-0.0228	0.0498	0.46	0.3525	0.2797	1.26
	-0.630	-0.6969	-0.0669	0.0518	1.29	0.3664	0.3360	1.09
	0.524	0.5379	0.0139	0.0224	0.62	0.1582	0.1601	0.99
	-0.375	-0.6405	-0.2655	0.1107	2.40	0.7828	0.4102	1.91
	0.8	0.7389	-0.0611	0.0716	0.85	0.5060	0.3121	1.62
	0.3	0.3548	0.0548	0.0568	0.96	0.4016	0.2716	1.48
70	1.031	1.0056	-0.0254	0.0439	0.58	0.3105	0.2115	1.47
	-0.630	-0.6225	0.0075	0.0449	0.17	0.3173	0.2535	1.25
	0.524	0.4974	-0.0266	0.0183	1.45	0.1294	0.1209	1.07
	-0.375	-0.5691	-0.1941	0.0753	2.58	0.5322	0.3120	1.71
	0.8	0.8253	0.0253	0.0538	0.47	0.3803	0.2352	1.62
	0.3	0.3256	0.0256	0.0460	0.56	0.3250	0.2049	1.59
100	1.031	0.9845	-0.0465	0.0264	1.76	0.1869	0.1770	1.06
	-0.630	-0.5846	0.0454	0.0307	1.48	0.2173	0.2120	1.03
	0.524	0.5061	-0.0179	0.0140	1.28	0.0991	0.1012	0.98
	-0.375	-0.5157	-0.1407	0.0485	2.90	0.3433	0.2578	1.33
	0.8	0.8483	0.0483	0.0335	1.44	0.2367	0.1967	1.20
	0.3	0.3185	0.0185	0.0327	0.57	0.2312	0.1713	1.35

Table 3.8

Summary Statistics for Model G

All Sample Sizes*

N	P	M	B	SE	B/SE	SD	ASD	SD/ASD
40	1.295	1.2377	-0.0573	0.0239	2.40	0.1691	0.1719	0.98
	-0.428	-0.4012	0.0268	0.0209	1.28	0.1475	0.1515	0.97
	1.0	1.1690	0.1690	0.0468	3.61	0.3310	0.3062	1.08
	0.5	0.5228	0.0228	0.0320	0.71	0.2262	0.1728	1.31
70	1.295	1.2674	-0.0276	0.0199	1.39	0.1406	0.1300	1.08
	-0.428	-0.4213	0.0067	0.0169	0.40	0.1199	0.1141	1.05
	1.0	1.1183	0.1183	0.0366	3.23	0.2585	0.2296	1.13
	0.5	0.5189	0.0189	0.0239	0.79	0.1687	0.1306	1.29
100	1.295	1.2697	-0.0253	0.0163	1.55	0.1154	0.1088	1.06
	-0.428	-0.4219	0.0061	0.0140	0.44	0.0993	0.0954	1.04
	1.0	1.1090	0.1090	0.0304	3.59	0.2149	0.1920	1.12
	0.5	0.5220	0.0220	0.0178	1.24	0.1260	0.1093	1.15

*Although Filter A(i) was used the variance ratio was set at 100/1 owing to the smaller roots in the A.R.

Tables 3.7 and 3.8 highlight the impact of an exogenous variable in that the biases evident in Model D are no longer present with Model F although the transfer functions are identical. Therefore it is likely that the results of Tables 3.2 - 3.4 set upper limits to the bias of an estimator and that the introduction of (positively) auto-correlated exogenous variables will improve the performance of the estimator.[6] However, whilst the A.R. and M.A. parameters are better estimated, that attached to the exogenous variable has high and significant biases in

6 As will become apparent later the root restrictions frequently have an effect that suggests superior sampling performance in samples of 40 than 70. This is particularly noticeable for the ratio SD/ASD so that the reader should concentrate on the behaviour of SD in isolation as well as its relation to the asymptotic value.

both models and in all sample sizes. In contrast to the bias the standard deviation of this parameter was not significantly different from its asymptotic value for the largest sample size in either model. Again an analysis of the trials yields the information that the root restriction was enforced 14, 4 and 1 times for N = 40, 70 and 100 for Model F and 2,0 and 0 for Model G. The Kolmogorov - Smirnov D statistic is tabulated for all sample sizes for Model G in Table 3.9.

Table 3.9

Kolmogorov - Smirnov D Statistic for Model G

All Sample Sizes

Model	Sample	D			
		β_1	β_2	γ_0	α_1
G	40	0.1863	0.1167	0.2484	0.1210
	70	0.1614	0.1078	0.1882	0.1638
	100	0.1562	0.0896	0.2004	0.1380

Referring to the confidence limits of Table 3.6 we see that β_1, β_2 and α_1 have D statistics that are compatible with normality when N = 100 but that for lower samples this is not the case. In all samples γ_0 seems to be non-normal and this is probably a consequence of the large bias noted in Table 3.8.

3.5 Misspecification and Lack of Identification

Misspecification is a universal problem in econometric analysis but it is accentuated in ARMA/ARMAX models owing to the ever present threat of a lack of identification. The two are related in the following manner. Assume that the model to be estimated is truly

ARMAX (p,r,q) but that an attempt is made to estimate it as ARMAX (p+1, r+1, q+1). Then by Condition 5 of Section 1.3 the resulting parameters are not identified. Several models were used to explore the problems arising from misspecification and/or lack of identification and five of these are reported on below.

MODEL H

$$(1-1.1L + 0.34L^2 + 0.12L^3)\, y(t) = (1+0.77L + 0.4L^2 + 0.05L^3)\, e(t)$$

MODEL I

$$(1-1.1L + 0.34L^2 + 0.12L^3)\, y(t) = (1+0.8L + 0.3L^2)\, e(t)$$

MODEL J

$$(1-1.031L + 0.630L^2 - 0.524L^3 + 0.0L^4)\, y(t) = -0.375\, x(t) + (1+0.8L + 0.6L^2)\, e(t)$$

MODEL K

$$(1-1.55L + 0.6L^2 + 0.0L^3)\, y(t) = (1+0.8L + 0.3L^2)\, e(t)$$

Appendix 8 contains the spectra associated with Models H and I and shows that the A.R. response function is quite complex. Table 3.10 reproduces the sampling statistics for Models H and I when the sample size is 100.

Table 3.10

Sampling Statistics for Models H and I

Sample Size N = 100

Model	P	M	B	SE	B/SE	SD	ASD	SD/ASD
H	1.1	1.3482	0.2482	0.0990	2.51	0.7001	18.52	0.04
	-0.34	-0.6614	-0.3214	0.1324	2.43	0.9361	23.94	0.04
	-0.12	0.0196	0.1396	0.0646	2.16	0.4569	11.02	0.04
	0.77	0.5207	-0.2493	0.1008	2.47	0.7126	18.53	0.04
	0.40	0.2395	-0.1605	0.0629	2.55	0.4448	10.73	0.04
	0.05	-0.0345	-0.0845	0.0386	2.19	0.2729	5.49	0.05
I	1.1	1.2460	0.1460	0.0716	2.04	0.5062	0.5637	0.90
	-0.34	-0.5393	-0.1993	0.0997	2.00	0.7051	0.8100	0.87
	-0.12	-0.0400	0.0800	0.0509	1.57	0.3596	0.4127	0.87
	0.8	0.6667	-0.1333	0.0748	1.78	0.5266	0.5507	0.96
	0.3	0.2097	-0.0903	0.0429	2.10	0.3036	0.2969	1.02

Table 3.10 is important for presenting a Model (Model H) that fulfils the identification conditions of Chapter 1 if these are interpreted very stringently i.e. identical common roots etc., but which might not be identified if only approximate conditions are required. A judgement on the source of any identification problem is complicated by the possibility that two conditions may have been partially violated - those relating to the location of zero elements in the transfer function and the presence of common roots - as the two final coefficients 0.12 and 0.05 do not differ greatly from zero but at the same time the roots of 0.2 in $B(L^{-1}) = 0$ and 0.1 in $A(L^{-1}) = 0$ are close. There does not seem to be any test for deciding which of these is a cause of a lack of identification. Of course if there was a complete lack of identification the likelihood would be flat and the existence of some curvature indicated that this was not the case. Notwithstanding this it was noticeable that substantial perturbations in the parameters caused only a small variation in the likelihood.

Actually Model H was responsible for a good deal of trouble as it was one of the first models examined and the poor performance shown in Table 3.10 (apparent even in samples of 200) was inexplicable until the notion of a mild form of non-identification was adopted.[7] An earlier report on this experiment (Pagan [88]) claimed that if the true values were used as starting values all biases were insignificant but the development of improved search routines and the construction of superior non-linear algorithms have invalidated this assertion. Nevertheless the earlier assertion is of interest as an example of the difficulties that may arise in the application of non-linear estimators to models with flat likelihoods and illustrates a danger of Monte Carlo experimentation with non-linear algorithms i.e. by chance initial values may be selected far from the true values and a flat likelihood may then result in a premature termination of the iterations. In such cases the summary statistics are adversely affected and experiments must be re-run with new starting values and/or new algorithms must be developed. Each experiment was run thrice - once with O.L.S. estimates, once with the true values, and once with arbitrary values as starting points. Such effort was generally wasteful in as much that many of these difficulties do not arise in empirical application or, if they do, because there is a need for only one trial, alternative methods of circumvention may be used e.g. multiple starting values.

Is it certain that the biases of Model H stem from mild non-identification rather than the nature of the model? No unequivocal reply

---

7 Asymptotically it is possible for such models to be perfectly identified as the range of possible values will become smaller.

can be given to this query but any verdict must consider the following facts.

(a) The response functions of Model H are similar to those of Model B (in that the squared gains of the A.R. response function are unimodal with the peak at roughly $\pi/5$) and as Table 3.5 shows no such biases were evident for the latter model.

(b) Model I, which has the same A.R. response function as Model H, also displays biases that are either marginally significant or insignificant but the asymptotic and empirical variances do not differ significantly. Model I has roots in both transfer functions that are quite distinct and non-zero final coefficients.

(c) The asymptotic covariances of Model H are quite large. If the model was not identified the likelihood would be flat, the second derivatives would be zero, and the Hessian would be singular so that one would expect that a mild lack of identification would lead to large entries in the covariance matrix.

Assuming that the above features are characteristic of an "unidentified" model little solace is to be gained from this example as the highest order coefficients do not seem unreasonable and there is no a priori belief that such parameters (and roots) are unlikely in modelling economic time series. Therefore it is desirable to seek practical means for recognizing any lack of identification. The two most frequently adopted centred on special characteristics of the likelihood function in such a situation.

(i) At least two sets of starting values were chosen. If the likelihood is flat radically different parameter sets will be associated

with a similar likelihood and if this was found there was an à fortiori case for lack of identification.

(ii) An index might be constructed by exploiting the relationship between the covariance matrix and identification i.e. large entries if the model is not identified. A more exact test lies in the magnitude of the ratio of the maximum to minimum eigenvalue of the weighting matrix - the most commonly used test of an ill-conditioned symmetric matrix (see Golub [32]).

Fortunately (ii) seems reliable. Two examples may be cited. Firstly Model H had an eigenvalue ratio of 10^7 versus 10^3 for all other asymptotic covariance matrices. Secondly a model of the form $(1-\beta_1 L)(1-\beta_4 L^4)\, y(t) = (1 + \alpha_1 L)\,(1 + \alpha_4 L^4)\, e(t)$ was fitted to the series "Imputed Rent on Dwellings" (one of the five components of Australian personal consumption expenditure modelled in Chapter 5) and the variances were so much higher than found by substituting the estimated parameters into asymptotic formulae as to suggest a possible lack of identification. Subsequently it was found that the series is constructed by interpolating annual data so that $\beta_4 = \alpha_4 = 0$ presenting a classic case of non-identification arising from misspecification. A spectral analysis confirmed this by disclosing the absence of seasonal peaks. The model was re-estimated successfully with the L^4 terms deleted.

Although (i) and (ii) are suitable tests, in another context they may be misleading viz. when there is multicollinearity between the regressors. If this is present it is well known that (ii) holds (Goldberger [30]) and in Monte Carlo experiments with such data H-T have

found that (i) occurs as well. To some extent multicollinearity may be regarded as an identification problem in that it implies an inability to statistically distinguish different estimates and in many cases the only solution would be to omit lagged y's - the solution given above for the imputed rent example. Incidentally this also invalidates a too hasty use of at least one important inferential method: that of "overfitting" the model (recommended by Box and Jenkins [9]) as one might be comparing identified and non-identified models thereby casting doubt upon the validity of any F of χ^2 statistics that would normally be formed.[8] Finally a simple way of distinguishing between multicollinearity and identification exists in ARMAX models as the addition of exogenous variables will aid identification but is unlikely to reduce multicollinearity.

Models J and K satisfy the identification requirements and thereby allow an exploration of the properties of the estimator under a misspecification. Only the A.R. is misspecified as H-T have dealt with misspecification of the order of the M.A. at length and earlier studies indicated that the results were similar for both situations. Tables 3.11 and 3.12 contain the sampling statistics for the two models.

---

[8] "Overfitting" is equivalent to a likelihood ratio test if performed properly. Obviously extending the order of all polynomials simultaneously will almost certainly eventually lead to a lack of identification and therefore this must be avoided.

Table 3.11

Sampling Statistics for Model J

All Sample Sizes

N	P	M	B	SE	B/SE	SD	ASD	SD/ASD
40	1.031	1.0804	0.0494	0.0729	0.68	0.5515	0.3411	1.62
	-0.630	-0.7946	-0.1646	0.1057	1.56	0.7474	0.5221	1.43
	0.524	0.6167	0.0927	0.0847	1.09	0.5987	0.4561	1.31
	0.0	-0.0614	-0.0614	0.0461	1.33	0.3260	0.2623	1.24
	-0.375	-0.7324	-0.3574	0.1348	2.65	0.9531	0.4345	2.19
	0.8	0.7981	-0.0019	0.0593	0.03	0.4192	0.3017	1.39
	0.6	0.6532	0.0532	0.0513	1.04	0.3624	0.1813	2.00
70	1.031	1.0630	0.0320	0.0536	0.60	0.3790	0.2575	1.47
	-0.630	-0.6855	-0.0555	0.0793	0.70	0.5610	0.3942	1.42
	0.524	0.5564	0.0324	0.0649	0.50	0.4596	0.3442	1.34
	0.0	-0.0527	-0.0527	0.0329	1.60	0.2325	0.1975	1.18
	-0.375	-0.5762	-0.2012	0.0757	2.66	0.5352	0.3386	1.58
	0.8	0.7762	-0.0238	0.0627	0.38	0.4392	0.2280	1.93
	0.6	0.6290	0.0290	0.0438	0.66	0.3101	0.1370	2.26
100	1.031	1.0367	0.0057	0.0421	0.14	0.2978	0.2154	1.38
	-0.630	-0.6576	-0.0276	0.0625	0.44	0.4417	0.3298	1.34
	0.524	0.5576	0.0336	0.0479	0.70	0.3384	0.2879	1.18
	0.0	-0.0515	-0.0515	0.0279	1.85	0.1974	0.1652	1.19
	-0.375	-0.5315	-0.1565	0.0535	2.93	0.3783	0.2822	1.34
	0.8	0.7899	-0.0101	0.0473	0.21	0.3345	0.1908	1.75
	0.6	0.6159	0.0159	0.0296	0.54	0.2091	0.1146	1.82

Table 3.12

Sampling Statistics for Model K

All Sample Sizes

N	P	M	B	SE	B/SE	SD	ASD	SD/ASD
40	1.55	1.4643	-0.0857	0.0673	1.27	0.4759	0.6872	0.69
	-0.6	-0.5566	0.0434	0.1085	0.40	0.7675	1.1859	0.65
	0.0	0.0203	0.0203	0.0481	0.42	0.3405	0.5407	1.63
	0.8	0.9055	0.1055	0.0719	1.47	0.5083	0.6678	0.76
	0.3	0.4245	0.1245	0.0495	2.52	0.3501	0.3926	0.89
70	1.55	1.6209	0.0709	0.0650	1.09	0.4593	0.5138	0.89
	-0.6	-0.7319	-0.1319	0.1089	1.21	0.7608	0.8866	0.86
	0.0	0.0692	0.0692	0.0479	1.44	0.3385	0.4042	0.84
	0.8	0.7478	-0.0522	0.0706	0.74	0.4994	0.4992	1.00
	0.3	0.2745	-0.0255	0.0508	0.50	0.3598	0.2935	1.23
100	1.55	1.6030	0.0530	0.0569	0.93	0.4024	0.4280	0.94
	-0.6	-0.7122	-0.1122	0.0951	1.18	0.6726	0.7385	0.91
	0.0	0.0490	0.0490	0.0426	1.15	0.3011	0.3367	0.89
	0.8	0.7545	-0.0455	0.0579	0.79	0.4099	0.4159	0.99
	0.3	0.2518	-0.0482	0.0415	1.16	0.2935	0.2446	1.20

Tables 3.11 and 3.12 show that all biases have risen from the correctly specified case (compare to Tables 3.2, 3.3 and 3.4 on pp 117-119 for Model K) but this is obscured by the complementary rise in standard deviations which has generally made biases insignificant: the most interesting example of this being that of the misspecified parameter and the sole exception being the exogenous variable.[9] In addition a comparison of the asymptotic standard deviations for Model K with those for Model E reveal that although the extra parameter is zero there has been a considerable inflation in these variances. The rise in standard deviations has been observed by other writers. Hendry and Trivedi argue that this is the principal effect while Hannan and Nicholls,

[9] One may validly argue that the correct standard deviations to judge the significance of bias are those of the correctly specified models and if this is done the conclusion will change for most parameters.

considering it from a theoretical viewpoint, have demonstrated that asymptotically the standard deviations of the estimated parameters will be larger than those for a correctly specified model (with an exception being the pure M.A. case when they are invariant). Another aspect of the variances is the slow decline in the empirical values tabulated above. This may be due to an inadequate number of replications but it is more likely that the root restrictions, which were applied in many more cases when the model was misspecified, are acting to reduce the variances of the parameters. By analogy with the theory of linear restrictions one would expect that a restricted estimator would have smaller variance than an unrestricted estimator and in the experiments on Models J and K restrictions are encountered more frequently in the smaller sample sizes.

These considerations emphasize the need for statistics to enable judgements of specification both prior and posterior to estimation. Box and Jenkins have propounded a number of these and in later chapters we will consider their use (and to some extent their usefulness).

3.6 Rational Lags and Expectations

Two experiments serve to illustrate the impact of rational lags.

MODEL L

$$(1-1.295L + 0.428L^2)\, y(t) = 1.0\, x(t) + (1-1.295L + 0.428L^2)\, e(t)$$

MODEL M

$$(1-1.55L + 0.6L^2)\, y(t) = 1.0\, x(t) + (1-1.55L + 0.6L^2)\, e(t)$$

x(t) was generated by scheme B(ii) for both Models. Tables 3.13 and 3.14 list the summary statistics for the Phillips estimator applied to the above models when the linear restrictions between the A.R. and M.A. parameters are employed and when they are ignored.

Table 3.13

Summary Statistics for Model L

All Sample Sizes*

N	P	M		B		SE		B/SE	
		R	U	R	U	R	U	R	U
40	1.295	1.2975	1.3006	0.0025	0.0056	0.0081	0.0073	0.31	0.77
	-0.428	-0.4293	-0.4318	-0.0013	-0.0038	0.0069	0.0063	0.19	0.60
	1.0	1.0005	0.9934	0.0005	-0.0066	0.0125	0.0114	0.04	0.58
	-1.295	-1.2975	-1.5129	-0.0025	-0.2179	0.0081	0.0256	0.31	8.51
	0.428	0.4293	0.5617	0.0013	0.1337	0.0069	0.0313	0.19	4.27
70	1.295	1.2961	1.2982	0.0011	0.0032	0.0043	0.0045	0.26	0.71
	-0.428	-0.4286	-0.4303	-0.0006	-0.0023	0.0037	0.0038	0.16	0.61
	1.0	0.9988	0.9953	-0.0012	-0.0047	0.0065	0.0068	0.18	0.69
	-1.295	-1.2961	-1.4502	-0.0011	-0.1552	0.0043	0.0265	0.26	5.86
	0.428	0.4286	0.5236	0.0006	0.0956	0.0037	0.0256	0.16	3.73
100	1.295	1.2952	1.2954	0.0002	0.0004	0.0032	0.0035	0.06	0.11
	-0.428	-0.4279	-0.4281	0.0001	-0.0001	0.0027	0.0030	0.04	0.03
	1.0	0.9973	0.9974	-0.0027	-0.0026	0.0046	0.0048	0.59	0.54
	-1.295	-1.2952	-1.3809	-0.0002	-0.0859	0.0032	0.0158	0.06	5.44
	0.428	0.4279	0.4781	-0.0001	0.0501	0.0027	0.0149	0.04	3.36

*The column headed R refers to the restricted estimator and that head U the unrestricted estimator.

Table 3.14

Summary Statistics for Model M

All Sample Sizes

T	P	M		B		SE		B/SE	
		R	U	R	U	R	U	R	U
40	1.55	1.5474	1.5585	-0.0026	0.0085	0.0088	0.0068	0.30	1.25
	-0.6	-0.5975	-0.6075	0.0025	-0.0075	0.0081	0.0062	0.31	1.21
	1.0	1.0182	0.9905	0.0182	-0.0095	0.0203	0.0160	0.90	0.59
	-1.55	-1.5474	-1.7176	0.0026	-0.1676	0.0088	0.0151	0.30	11.10
	0.6	0.5975	0.7244	-0.0025	0.1244	0.0081	0.0170	0.31	7.32
70	1.55	1.5521	1.5557	0.0021	0.0057	0.0044	0.0041	0.48	1.39
	-0.6	-0.6019	-0.6054	-0.0019	-0.0054	0.0040	0.0037	0.48	1.46
	1.0	1.0018	0.9977	0.0018	-0.0023	0.0095	0.0098	0.19	0.23
	-1.55	-1.5521	-1.6903	-0.0021	-0.1403	0.0044	0.0200	0.48	7.02
	0.6	0.6019	0.7051	0.0019	0.1051	0.0040	0.0208	0.48	5.05
100	1.55	1.5545	1.5531	0.0045	0.0031	0.0038	0.0036	1.18	0.86
	-0.6	-0.6041	-0.6028	-0.0041	-0.0028	0.0034	0.0033	1.21	0.85
	1.0	0.9915	0.9938	-0.0085	-0.0062	0.0085	0.0076	1.00	0.82
	-1.55	-1.5545	-1.6504	-0.0045	-0.1004	0.0038	0.0154	1.18	6.52
	0.6	0.6041	0.6805	0.0041	0.0805	0.0034	0.0150	1.21	5.37

Received theory would suggest that asymptotically the principal difference between the estimators will be one of efficiency. It is therefore of some consequence that a perusal of Tables 3.13 and 3.14 identifies three differences in small samples.

(i) No clear pattern emerges regarding the biases of each estimator from a consideration of the A.R. parameters but the restricted estimates of the M.A. parameters exhibit smaller bias in all samples.

(ii) The bias in the M.A. parameters of the unrestricted estimator is likely to be a function of the size of the roots of the M.A. polynomial and this bias is quite large even in samples of 100.

(iii) The expected efficiency ranking has eventuated only for the M.A. parameters in all samples and the A.R. parameters in Model L for the two larger samples. There can be no certain explanation of this but two factors must be considered:

(a) The theorem on efficiency rankings pertains to two unbiassed estimators and it is not possible to prove that this is the case for the above estimators. Certainly some of the parameters are seveiely biassed.

(b) There may be a second restriction operative in the case of the "unrestricted" estimator in the form of the root restriction. Table 3.15 shows the number of trials for which the root restrictions were applied.

Table 3.15

Frequency of Usage of Root Restrictions

Models L and M: All Sample Sizes

Model	Estimator	40	70	100
L	R	0	0	0
	U	42	21	8
M	R	0	0	0
	U	45	34	17

We note that the restrictions are heavily used for the second estimator in both models with the number of applications being close for each model in the smallest sample. From this one might argue that, if the imposition of the restraints is acting to reduce the variances, the most likely sample size for Model L at which there will be perverse behaviour is N = 40. Table 3.13 supports the hypothesis.

Another view of the problem may be gained by a comparison of the asymptotic and sample variances. Table 3.16 contains the usual test statistic relating to this.

Table 3.16

Comparison of Asymptotic and Sample Standard Deviations

Models L and M: All Sample Sizes

Model	N	P	R			U		
			SD	ASD	SD/ASD	SD	ASD	SD/ASD
L	40	AR	0.0571	0.0329	1.74	0.0517	0.0381	1.36
		AR	0.0491	0.0314	1.56	0.0441	0.0321	1.37
		EV	0.0887	0.0527	1.68	0.0806	0.0538	1.50
		MA	0.0571	0.0329	1.74	0.1808	0.1515	1.19
		MA	0.0491	0.0314	1.56	0.2215	0.1509	1.47
	70	AR	0.0307	0.0256	1.20	0.0315	0.0290	1.09
		AR	0.0262	0.0240	1.09	0.0266	0.0244	1.09
		EV	0.0457	0.0405	1.13	0.0480	0.0410	1.17
		MA	0.0307	0.0256	1.20	0.1877	0.1134	1.66
		MA	0.0262	0.0240	1.09	0.1809	0.1123	1.61
	100	AR	0.0224	0.0239	0.94	0.0247	0.0242	1.02
		AR	0.0191	0.0201	0.95	0.0211	0.0204	1.03
		EV	0.0315	0.0339	0.93	0.0338	0.0343	0.99
		MA	0.0224	0.0239	0.94	0.1115	0.0944	1.18
		MA	0.0191	0.0201	0.95	0.1054	0.0935	1.13
M	40	AR	0.0623	0.0304	2.05	0.0478	0.0310	1.54
		AR	0.0575	0.0276	2.08	0.0493	0.0282	1.75
		EV	0.1435	0.0653	2.20	0.1129	0.0665	1.70
		MA	0.0623	0.0304	2.05	0.1066	0.1335	0.80
		MA	0.0575	0.0276	2.08	0.1199	0.1329	0.90
	70	AR	0.0312	0.0244	1.28	0.0290	0.0247	1.17
		AR	0.0285	0.0222	1.28	0.0262	0.0225	1.16
		EV	0.0673	0.0534	1.26	0.0693	0.0540	1.28
		MA	0.0312	0.0244	1.28	0.1416	0.1001	1.41
		MA	0.0285	0.0222	1.28	0.1471	0.0996	1.48
	100	AR	0.0268	0.0206	1.30	0.0253	0.0207	1.22
		AR	0.0244	0.0187	1.30	0.0232	0.0188	1.23
		EV	0.0602	0.0449	1.34	0.0536	0.0453	1.18
		MA	0.0268	0.0206	1.30	0.1092	0.0834	1.31
		MA	0.0244	0.0187	1.30	0.1058	0.0830	1.27

Table 3.16 shows that the asymptotic variances of the restricted

estimator (of the A.R. parameters) are always smaller than the unrestricted but by a very small amount. It is not surprising then that the empirical variances may have a ranking that is the reverse of the asymptotic one if an additional assumption is imposed upon the estimators.

These experiments lead to the conclusion that the discrepancies between the restricted and unrestricted estimators of the A.R. parameters will not be large in any sample size but that for the M.A. parameters considerable differences exist in both means and variances. Moreover the biases are such that estimation without restrictions will provide some idea of the approximate magnitude of the parameters and, if it is easiest to do this, should be done first. Subsequently substantial efficiency gains for the M.A. parameters are possible if the restrictions are appropriate.

3.7 Seasonal Models

As mentioned earlier in the chapter some experiments would be conducted to assess the impact of a seasonal pattern in x(t) and/or the disturbances. Three models were selected to reflect the type of equation that might arise with quarterly data.

MODEL N

$$(1-1.295L + 0.428L^2)\, y(t) = 1.0\, x(t) + (1+0.6L)(1+0.3L^4)\, e(t)$$

MODEL G/S

$$(1-1.295L + 0.428L^2)\, y(t) = 1.0\, x(t) + (1 + 0.5L)\, e(t)$$

MODEL O

$$(1-0.6L)^3\, y(t) = (1-0.6)^3\, x(t) + (1 + 0.3L)(1-0.7L^4)\, e(t)$$

The second of these models should be familiar as Model G of section 3.4 and "/S" distinguishes it as one with a seasonal pattern in the generated data. This pattern was achieved in the three models by usage of the following filters and variance ratios.

(i) Scheme A(i) with variance ratios of 500/1 and 100/1 for Models N and O respectively.

(ii) Scheme A with $\delta_1 = 0.99$, $\delta_2 = 0.8$, $\delta_3 = 0.8$, $\delta_4 = 0$ and variance ratio of 500/1 for Model G/S.

From Appendix 8 the spectra of y(t) in Models N and O and x(t) in G/S have a peak at $\pi/2$ and the lag distribution associated with Model O is unimodal with the peak response at four quarters.

Table 3.17 embodies the summary statistics for Models N and G/S at all sample sizes and Table 3.18 presents these for Model O when $N = 100$.

Table 3.17

Summary Statistics for Models N and G/S

All Sample Sizes

Model	N	P	M	B	SE	B/SE	SD	ASD	SD/ASD
N	40	1.295	1.2414	-0.0536	0.0258	2.08	0.1825	0.1719	1.06
		-0.428	-0.4410	-0.0130	0.0297	0.44	0.2103	0.1644	1.28
		1.0	1.2697	0.2697	0.0964	2.80	0.6819	0.4910	1.39
		0.6	0.6108	0.0108	0.0319	0.34	0.2259	0.1539	1.47
		0.3	0.4516	0.1516	0.0432	3.51	0.3056	0.1322	2.31
	70	1.295	1.2720	-0.0230	0.0186	1.24	0.1318	0.1299	1.01
		-0.428	-0.4399	-0.0119	0.0184	0.65	0.1299	0.1240	1.05
		1.0	1.1061	0.1061	0.0607	1.75	0.4290	0.3793	1.13
		0.6	0.6276	0.0276	0.0201	1.37	0.1425	0.1163	1.23
		0.3	0.3134	0.0134	0.0204	0.66	0.1440	0.0999	1.44
	100	1.295	1.2742	-0.0208	0.0147	1.41	0.1042	0.1087	0.96
		-0.428	-0.4362	-0.0082	0.0136	0.60	0.0961	0.1037	0.93
		1.0	1.0872	0.0872	0.0479	1.82	0.3391	0.3192	1.06
		0.6	0.6209	0.0209	0.0148	1.41	0.1044	0.0973	1.07
		0.3	0.2999	-0.0001	0.0148	0.01	0.1043	0.0836	1.25
G/S	40	1.295	1.2414	-0.0536	0.0257	2.09	0.1819	0.1719	1.06
		-0.428	-0.4338	-0.0058	0.0235	0.25	0.1663	0.1515	1.10
		1.0	1.4439	0.4439	0.0829	5.35	0.5861	0.3062	1.91
		0.5	0.5051	0.0051	0.0306	0.17	0.2161	0.1728	1.25
	70	1.295	1.2611	-0.0339	0.0214	1.58	0.1513	0.1300	1.16
		-0.428	-0.4258	0.0022	0.0206	0.11	0.1453	0.1141	1.27
		1.0	1.1993	0.1993	0.0499	3.99	0.3528	0.2296	1.54
		0.5	0.5250	0.0250	0.0220	1.14	0.1556	0.1306	1.19
	100	1.295	1.2635	-0.0315	0.0182	1.73	0.1290	0.1088	1.19
		-0.428	-0.4220	0.0060	0.0165	0.36	0.1166	0.0954	1.22
		1.0	1.1509	0.1509	0.0392	3.85	0.2775	0.1920	1.45
		0.5	0.5286	0.0286	0.0187	1.53	0.1334	0.1093	1.22

Table 3.18

Sampling Statistics for Model O

Sample Size N = 100

P	M	B	SE	B/SE	SD
0.6	0.5923	-0.0077	0.0059	1.31	0.0417
0.3	0.3257	0.0257	0.0203	1.27	0.1435
-0.7	-0.7886	-0.0886	0.0153	5.79	0.1082

The results for both models N and G/S are reassuring. There is a hint of substantial biases in the small sample of 40 but with the exception of γ_0 in Model G/S a sample of size 70 should be adequate for the application of asymptotic theorems. Of greatest interest in the Table is the smaller bias in γ_0 (and a variance closer to the asymptotic value) in Model N than in G/S: Model N being the only case encountered where the bias of γ_0 was insignificant when using the Nerlove filter. Model G/NS (p. 148) has the same transfer functions as G/S but differs in that Scheme A(ii) was used to produce non-stationarity in x(t); G/NS exhibits a substantial bias for γ_0. All in all however the transition to time series that have seasonal patterns does not seem to have affected the sampling performance of the estimator to any great extent, and the conclusions reached in the preceding sections should be applicable to a variety of time series.

Table 3.18 furnishes a bleaker picture. Only the statistics for N = 100 are reported owing to the large bias in α_4. As this was associated with a tendency to seek a minimum to the sum of squares outside the boundary (in this case $\alpha_4 = 1$), and later it will be seen that the frequency of such behaviour is inversely related to sample size, there did not seem to be any benefits arising from an examination of a smaller sample. The replications for which the restriction was enforced were 9,10,14,22,25,26 and 38. To further explore this the median was computed to be -0.7707, the range -0.5809 to -0.999, and the frequency distribution as in Table 3.19.

Table 3.19

Frequency Distribution for $\hat{\alpha}_4$ of Model 0.

Class	Frequency
$-\infty < x \leq 0.6$	0.02
$0.6 < x \leq 0.65$	0.04
$0.65 < x \leq 0.7$	0.14
$0.7 < x \leq 0.75$	0.20
$0.75 < x \leq 0.8$	0.18
$0.8 < x \leq 0.85$	0.16
$0.85 < x \leq 0.9$	0.12
$0.9 < x \leq 0.95$	0.0
$0.95 < x < 1$	0.14

Table 3.19 epitomizes the skewed distributions found in experiments in which the root restrictions were enforced a large number of times. In fact the skewness is even more accentuated than Table 3.19 suggests if one realizes that the 7 observations in the last class all lie between 0.99 and 1 so that the distribution is certainly non-normal. Further discussion on this point is deferred until similar experiments in the following chapter but it is important to note that:-

(i) Neither α_1 nor β_1 are significantly biassed in this sample and the distributions for these are more akin to the parameters of Models N and G/S.

(ii) The introduction of an exogenous variable may reduce the probability of such an occurrence e.g. the roots in the M.A. of Model N do not differ greatly in modulus from those of Model T in the next chapter, yet when N = 40 the boundary is encountered in a large number of replications for the latter (see Table 4.6 on p 171).

3.8 Stationarity

Stationarity has two aspects:-

(i) x(t) may be stationary but y(t) non-stationary owing to the presence of unit roots in B(L). If these roots are known to be unity, then by appropriate differencing of y(t) the resulting series will be stationary and the conclusions cited earlier will apply. H-T have explored the case when (unknown) unit roots are present in B(L) and have reached substantially the same conclusions as in situations where the roots lie strictly outside the unit circle (although as one would expect the biases must be larger if the restrictions are enforced).

(ii) x(t) may be non-stationary. As explained in 2.8 the theoretical distributions of ARMAX estimators are unknown but, relying on the analogy with classical regression, one is inclined to allow x(t) to follow any pattern that conforms to Grenander's conditions and an investigation will be made into the situation when x(t) has a polynomial trend.

To explore this effect three models are selected.

MODEL P

$$(1-0.6L)^3\, y(t) = (1-0.6)^3\, x(t) + (1+0.4L)\, e(t)$$

MODEL G/NS

$$(1-1.295L + 0.428L^2)\, y(t) = 1.0\, x(t) + (1 + 0.5L)\, e(t)$$

MODEL F/NS

$$(1-1.031L + 0.630L^2 - 0.524L^3)\, y(t) = -0.375\, x(t) + (1+0.8L+0.3L)^2\, e(t).$$

The exogenous variables are constructed in the following manner:

(i) Scheme A(ii) with variance ratios of 100/1 and 500/1 for Models P and G/NS.

(ii) Scheme A with $\delta_1 = 1.0$, $\delta_2 = 1.0$, $\delta_3 = 0.3$, $\delta_4 = -0.7$ and variance ratio of 500/1 for F/NS.

Table 3.20 sets out the statistics for P and G/NS and Table 3.22 presents these for F/NS.

Table 3.20

Summary Statistics for Models P and G/NS

All Sample Sizes

Model	N	P	M	B	SE	B/SE	SD
P	40	0.6	0.5821	-0.0179	0.0065	2.75	0.0463
		0.4	0.4442	0.0442	0.0262	1.69	0.1859
	70	0.6	0.5892	-0.0108	0.0054	2.00	0.0383
		0.4	0.4400	0.0400	0.0194	2.06	0.1375
	100	0.6	0.5891	-0.0109	0.0050	2.18	0.0357
		0.4	0.4350	0.0350	0.0163	2.15	0.1149
G/NS	40	1.295	1.2684	-0.0266	0.0228	1.17	0.1613
		-0.428	-0.4535	-0.0255	0.0234	1.09	0.1658
		1.0	1.3635	0.3635	0.0892	4.08	0.6305
		0.5	0.4990	-0.0010	0.0358	0.03	0.2534
	70	1.295	1.2698	-0.0252	0.0209	1.21	0.1476
		-0.428	-0.4314	-0.0034	0.0215	0.16	0.1520
		1.0	1.1721	0.1721	0.0499	3.45	0.3526
		0.5	0.5221	0.0221	0.0214	1.03	0.1513
	100	1.295	1.2728	-0.0222	0.0184	1.21	0.1301
		-0.428	-0.4290	-0.0010	0.0180	0.06	0.1276
		1.0	1.1449	0.1449	0.0414	3.50	0.2929
		0.5	0.5260	0.0260	0.0206	1.26	0.1457

Table 3.21

Summary Statistics for Model F/NS

All Sample Sizes

N	P	M	B	SE	B/SE	SD
40	1.031	0.8956	-0.1354	0.0339	3.99	0.2398
	-0.630	-0.6042	0.0258	0.0383	0.67	0.2708
	0.524	0.4813	-0.0427	0.0297	1.44	0.2101
	-0.375	-1.0719	-0.6969	0.1272	5.48	0.8997
	0.8	0.8943	0.0943	0.0428	2.20	0.3028
	0.3	0.4583	0.1583	0.0442	3.58	0.3129
70	1.031	0.9337	-0.0973	0.0338	2.88	0.2387
	-0.630	-0.5727	0.0573	0.0378	1.52	0.2672
	0.524	0.4611	-0.0629	0.0188	3.35	0.1333
	-0.375	-0.8566	-0.4816	0.0677	7.11	0.4786
	0.8	0.8740	0.0740	0.0399	1.85	0.2821
	0.3	0.3756	0.0756	0.0350	2.16	0.2473
100	1.031	0.9450	-0.0860	0.0226	3.81	0.1598
	-0.630	-0.5622	0.0678	0.0281	2.41	0.1990
	0.524	0.4717	-0.0523	0.0141	3.71	0.0998
	-0.375	-0.7038	-0.3288	0.0439	7.49	0.3106
	0.8	0.8795	0.0795	0.0284	2.80	0.1989
	0.3	0.3455	0.0455	0.0275	1.65	0.1946

Tables 3.20 and 3.21 provide conflicting evidence on the impact of non-stationarity in x(t) upon the distribution of the estimators. Both Models P and F/NS have substantial biases in parameters in most sample sizes whereas Model G/NS does so only for γ_0. Because of this there must be some doubt about the performance of the estimator under what may be regarded as realistic conditions. Certainly when contrasted with Model F (p 126), the statistics for F/NS indicate that non-stationarity in the exogenous variable may have a very marked effect upon the estimator.

3.9 Departures from Normality

In this final section we are concerned with the sampling properties of the Phillips estimator under the assumption that the disturbances

are i.i.d. rather than n.i.d. To achieve this objective series of uniformly distributed random numbers were substituted for normally distributed numbers and the parameters of the uniform distribution were set to yield the same means and variances as in the Gaussian experiments. The models selected for study are Models B and F and the summary statistics are contained in Table 3.22. The relevant comparison is with Tables 3.2, 3.3, 3.4 and 3.6(p.117-199 and 126).

Table 3.22

Summary Statistics for Models B and F.

All Samples Sizes: Disturbances i.i.d.

Model	N	P	M	B	SE	B/SE	SD	ASD	SD/ASD
B	40	1.3	1.2982	-0.0018	0.0302	0.06	0.2137	0.1901	1.12
		-0.6	-0.5913	0.0087	0.0283	0.31	0.2002	0.1759	1.14
		0.6	0.6563	0.0563	0.0426	1.32	0.3016	0.2183	1.38
		0.3	0.3187	0.0187	0.0479	0.39	0.3392	0.1991	1.70
	70	1.3	1.3120	0.0120	0.0198	0.61	0.1400	0.1437	0.97
		-0.6	-0.6045	-0.0045	0.0164	0.27	0.1165	0.1329	0.88
		0.6	0.5991	-0.0009	0.0207	0.04	0.1985	0.1650	1.20
		0.3	0.2876	-0.0124	0.0273	0.45	0.1929	0.1505	1.28
	100	1.3	1.2920	-0.0080	0.0165	0.48	0.1172	0.1202	0.98
		-0.6	-0.5893	0.0107	0.0153	0.70	0.1078	0.1112	0.97
		0.6	0.6266	0.0266	0.0250	1.06	0.1769	0.1380	1.28
		0.3	0.3084	0.0084	0.0210	0.40	0.1485	0.1259	1.18
F	40	1.031	0.9664	-0.0646	0.0585	1.10	0.4134	0.2797	1.48
		-0.630	-0.5646	0.0654	0.0536	1.22	0.3792	0.3360	1.13
		0.524	0.4114	-0.1126	0.0296	3.80	0.2090	0.1601	1.31
		-0.375	-0.4581	-0.0831	0.1839	0.45	1.3002	0.4102	3.17
		0.8	0.8736	0.0736	0.0700	1.05	0.4953	0.3121	1.59
		0.3	0.3231	0.0231	0.0690	0.33	0.4877	0.2716	1.80
	70	1.031	0.9510	-0.0800	0.0421	1.90	0.2976	0.2115	1.41
		-0.630	-0.5467	0.0833	0.0478	1.74	0.3381	0.2535	1.33
		0.524	0.4705	-0.0535	0.0210	2.55	0.1485	0.1209	1.23
		-0.375	-0.4373	-0.0623	0.0883	0.71	0.6241	0.3120	2.00
		0.8	0.8974	0.0974	0.0469	2.08	0.3317	0.2352	1.41
		0.3	0.3556	0.0556	0.0472	1.18	0.3337	0.2049	1.63
	100	1.031	0.9973	-0.0337	0.0292	1.15	0.2064	0.1770	1.17
		-0.630	-0.6100	0.0200	0.0378	0.53	0.2674	0.2120	1.26
		0.524	0.5082	-0.0158	0.0194	0.81	0.1374	0.1012	1.36
		-0.375	-0.4077	-0.0327	0.0459	0.71	0.3247	0.2578	1.26
		0.8	0.8504	0.0504	0.0329	1.53	0.2327	0.1967	1.18
		0.3	0.3267	0.0267	0.0312	0.86	0.2207	0.1713	1.29

A comparison of the various tables reveals that the estimator performs in a similar fashion when the disturbances are i.i.d. rather than n.i.d. Perhaps the most obvious differences are the smaller (and insignificant) bias of γ_0 and the tendency towards larger variances in the ARMAX model when disturbances are i.i.d. Ranged against this must be the larger biases of β_1, β_2, and β_3 in this model under the more general disturbance format suggesting the speculative conclusion that the estimator may perform differently for individual parameters but similarly if a macroscopic view is taken of the experiments. When the hypothesis that the means and variances of both estimators were not significantly different was tested it was accepted at the 5% level for all parameters when N = 100.[10]

3.10 Conclusion

Chapter 3 has highlighted the distributional properties of the Phillips estimator in small to large samples. Most of the investigation has been concerned with a comparison of the sampling and asymptotic means and variances but some attention has been paid to the other moments.

It has been concluded that the Phillips estimator is fairly robust against a wide variety of model specification and data characteristics - although some reservation has been expressed concerning the sample size required for asymptotic theorems to hold when there is non-stationarity in the exogenous variable. If this case is disregarded it seemed that samples of 70-100 would be sufficient for these theorems to be used.

---

10 For smaller samples normality is unlikely and the test could be misleading.

Appendix 7

A Critique of the Hendry and Trivedi Study

To date the main source of information on the small sample properties of the Phillips estimator is the lengthy report issued by D.F. Hendry and P.K. Trivedi (H-T) [49]. This report is a condensation of a number of earlier papers by the authors. It features a wide range of parameter sets for the A.R. and M.A. transfer functions and investigates a number of factors such as the presence of (unknown) unit roots in B(L) and multicollinearity between regressors in a model resembling a neo-classical investment function. As such it is of great value to economists who are involved in the estimation of ARMAX models, but its utility is to some extent marred by a number of unsatisfactory aspects that deserve comment.

The first of these arises from statements by the authors about the relationship between the number of trials performed and the number retained for use. They say

> "Individual runs which yielded ridiculous parameter estimates were rejected" [49, p.10].

This admission highlights a crucial problem that is common to all Monte Carlo work i.e. under what circumstances is the deletion of results from any replication justified? If the argument is advanced that these estimates violate the restrictions imposed by Conditions 4-7 of Chapter 1 then it is only proper that such beliefs should be incorporated into the estimation procedure rather than to impose them by mechanistic rejection, as adherence to the latter method will inevitably bias the estimates computed from the retained set of numbers. In the simple case when the

offending polynomial is only of first order the rejected values would lie in the tails of the distribution so that there would be a bias towards zero but in other cases, although the direction of bias of the roots may be established, it is difficult to relate this to the parameters.

As H-T do not state their criteria for such judgements it is impossible to evaluate exactly the extent and direction of bias (if any) that is contained in their report. For the moment it is sufficient to record the attitude that the dangers inherent in a subjective evaluation of the worth of any individual estimate outweigh any gain from discarding it and for this reason all trials were utilized in the calculation of the statistics reported in Chapter 3.

H-T propose a regression model as a convenient mode for the analysis of the sampling output. To give an example, in Part I of their paper the estimated biases of the parameters $(\bar{\theta}-\theta)$ are explained by a linear combination of a constant, the true value of the M.A. parameters (α_1,α_2), and two dummy variables Z_1 and Z_2 which are employed to account for sample size and the nature of the exogenous variable respectively. This equation is

$$\bar{\theta} - \theta = c_0 + c_1\alpha_1 + c_2\alpha_2 + c_3Z_1 + c_4Z_2 + v \qquad \text{(A7.1)}$$

where $E(v) = 0 \quad E(vv^T) = \sigma_v^2$

$Z_1 = 0$ if N = 25
$= 1$ if N = 50

$Z_2 = 0$ if x(t) was fixed
$= 1$ if x(t) is stochastic.

For each experiment values of α_1, α_2, Z_1 and Z_2 are available and a regression on this data is designed to isolate the effects of these influences upon the bias vector $(\bar{\theta}-\theta)$.

Such an approach is suggestive as the mass of data produced in Monte Carlo experiments is frequently difficult to digest and this allows the salient features of any group of experiments to be grasped within a familiar framework. However although the method has great potential it is also true that care must be taken in the specification of (A7.1) if one is to interpret the regression parameters in a useful way, and it is at this stage that H-T appear to have erred. Briefly the error originates in the unweighted measure of bias adopted. This is obviously inappropriate in a stochastic context in which the absolute size of a variable should always be related to its standard deviation. Therefore in (A7.1) it is not $(\bar{\theta}-\theta)$ that should be the dependent variable but rather $(\bar{\theta}-\theta)/\sigma_{(\bar{\theta}-\theta)}$ i.e. the ratio of the bias to the standard error of the bias. Recognizing this we note that even though the magnitude of the bias may vary with the experiment this variation may not be significant. The asymptotic covariance matrix derived in Appendix 4 shows that the standard error of the bias will be a function of the true values of all parameters in the model as well as the signal/noise ratio so that it is to be expected that the coefficients attached to these explanators in any regression would be significant. However the chain of causation is such as to make the result of little interest.

The above discussion provides a partial answer to H-T's puzzlement over the significance of c_1 and c_2 in their regressions of type (A7.1).[1] There are other regressions in the paper which may also be criticized

---

[1] In an early draft of Part II of their paper they stated "One would like to explain why the coefficients c_1 and c_2 are significantly different from zero even when the maximum likelihood estimators are used."

for the type of variable included. An example lies in the variable representing the signal/noise ratio which is a curious amalgam of variances and M.A. parameters produced by the use of σ_u^2/σ_e^2 where $u(t) = A(L)\ e(t)$. This specification introduces a correlation between the variables of the regression which is entirely unnecessary and thereby reduces the extent to which $(c_0,\ldots,c_4)$ may be relied upon as isolating the separate influences of each force i.e. as far as possible good experimental design will require orthogonality of the regressors.

Finally it is conceivable that some of the explanation of bias may reside in the starting values employed by H-T. If one begins by setting the M.A. parameters to zero (and H-T appear to have adopted this strategy for some of their work) it is possible that the presence of local minima and/or extreme flatness of the likelihood will result in a convergence to estimates that are biassed toward zero, and furthermore it is likely that the higher the absolute value of the M.A. parameters the more pronounced will this tendency be. If this is the case then the dependence of the bias upon the level of α_1 and α_2 might be explained by a failure to converge to M.L. estimates. For at least one experiment with collinear data the authors report multiple maxima and flat likelihoods.

To recapitulate the above arguments there seems some doubt concerning the mode chosen to explain bias. Certainly the inclusion of α_1, α_2 and possibly the signal/noise ratio must be regarded with scepticism as the old adage "correlation does not imply causation" is highly relevant in this context. As the above discussion suggests there may be a number of causes which operate through the variables of

the regression and unless one identifies the primary causes little of value will be gleaned from the exercise.[2] On the other hand the individual regression results are very useful, and apart from the caveat concerning the number of retained trials cannot be criticized.

---

2 This conclusion would be too strong if variables such as sample size and the nature of the regressors explained a large proportion of the variation in the bias. A cursory examination of the t-statistics for the regressions reveals that from 30-90% of the variation is accounted for by α_1 and α_2 and the signal/noise ratio (the upper limit being reached for the experiments of Part 3). Of course as we have mentioned above the possibility of non-orthogonality between the regressors must make such a claim inexact.

Appendix 8

Characteristics of Models A-U

This appendix serves as a guide to the models used in the simulations of Chapters 3 and 4. Tables A8.1 and A8.2 contain the roots and transfer functions of all models. Following this are graphs of

(i) The inverse of the squared gain of the A.R. i.e. $1/|B(\lambda)|^2$.

(ii) The squared gain of the M.A. i.e. $|A(\lambda)|^2$.

(iii) The spectrum of x(t) (if the model is ARMAX) i.e. $f_{xx}(\lambda)$.

(iv) The spectrum of y(t) i.e. $f_{yy}(\lambda)$.

If the model is a pure M.A. (e.g. Models T and U) only $f_{yy}(\lambda)$ is listed, and if ARMA, but not ARMAX, (i), (ii) and (iv) are presented. When the model is ARMAX it was customary to plot the logarithm (to base 10) of (i), (iii) and(iv) so that the scales vary between graphs. Additionally, in order to avoid computational difficulties, filters such as $(1-L)(1-L^4)$ were replaced by $(1-0.99L)(1-0.99L^4)$ in the construction of $f_{xx}(\lambda)$. Although an approximation, the procedure should provide a correct visual impression. Finally the weighting functions describing the lag distribution of selected ARMAX models are presented.

The parameters of Table A8.1 apply to the ARMAX model

$$B(L)\, y(t) = C(L)\, x(t) + A(L)\, e(t),$$

with the exogenous variables generated by

$$D(L)\, x(t) = F(L)\, \varepsilon(t).$$

The column headed V refers to the variance ratio $\sigma_e^2/\sigma_\varepsilon^2$.

Table A8.1

Transfer Functions of Models A-U

Model	B(L)	A(L)	C(L)	D(L)	F(L)	V
A	$1-0.8L$	$1+0.5L$				
B	$1-1.3L+0.6L^2$	$1+0.6L+0.3L^2$				
C	$1-1.57L+1.23L^2-0.42L^3$	$1+1.2L+0.62L^2+0.27L^3$				
D	$1-1.031L+0.630L^2-0.524L^3$	$1+0.8L+0.3L^2$				
E	$1-1.55L+0.6L^2$	$1+0.8L+0.3L^2$				
F	$1-1.031L+0.630L^2-0.524L^3$	$1+0.8L+0.03L^2$	-0.375	$(1-0.95L)(1-0.75L)$	$1+0.8L$	500
G	$1-1.295L+0.428L^2$	$1+0.5L$	1.0	$(1-0.95L)(1-0.75L)$	$1+0.8L$	100
H	$1-1.1L+0.34L^2+0.12L^3$	$1+0.77L+0.4L^2+0.05L^3$				
I	$1-1.1L+0.34L^2+0.12L^3$	$1+0.8L+0.3L^2$				
J	$1-1.031L+0.630L^2-0.524L^3+0.0L^4$	$1+0.8L+0.6L^2$	-0.375	$(1-0.95L)(1-0.75L)$	$1+0.8L$	500
K	$1-1.55L+0.6L^2+0.0L^3$	$1+0.8L+0.3L^2$				
L	$1-1.295L+0.428L^2$	$1-1.295L+0.428L^2$	1.0	$1-0.9L$		1
M	$1-1.55L+0.6L^2$	$1-1.55L+0.6L^2$	1.0	$1-0.9L$		10
N	$1-1.295L+0.428L^2$	$(1+0.6L)(1+0.3L^4)$	1.0	$(1-0.95L)(1-0.75L)$	$1+0.8L$	500
G/S	$1-1.295L+0.428L^2$	$1+0.5L$	1.0	$(1-0.99L)(1-0.8L^4)$	$1+0.8L$	500
O	$(1-0.6L)^3$	$(1+0.3L)(1-0.7L^4)$	$(1-0.6)^3$	$(1-0.95L)(1-0.75L)$	$1+0.8L$	100
P	$(1-0.6L)^3$	$1+0.4L$	$(1-0.6)^3$	$(1-L)(1-L^4)$	$(1-0.082L)(1-0.372L^4)$	100
G/NS	$1-1.295L+0.428L^2$	$1+0.5L$	1.0	$(1-L)(1-L^4)$	$(1-0.082L)(1-0.372L^4)$	500
F/NS	$1-1.031L+0.630L^2-0.524L^3$	$1+0.8L+0.3L^2$	-0.375	$(1-L)(1-L^4)$	$(1+0.3L)(1-0.7L^4)$	500
Q	$1-0.8L$	$1+0.5L$	-0.3	$(1-0.95L)(1-0.75L)$	$1+0.8L$	100
R	$1-1.031L+0.630L^2-0.524L^3$	$1+0.5L$	-0.375	$(1-0.95L)(1-0.75L)$	$1+0.8L$	500
S	$1-1.295L+0.428L^2$	$1+0.5L$	1.0	$1-0.9L$		1
T		$(1+0.2L)(1+0.4L^4)$				
U		$1+0.5L$				

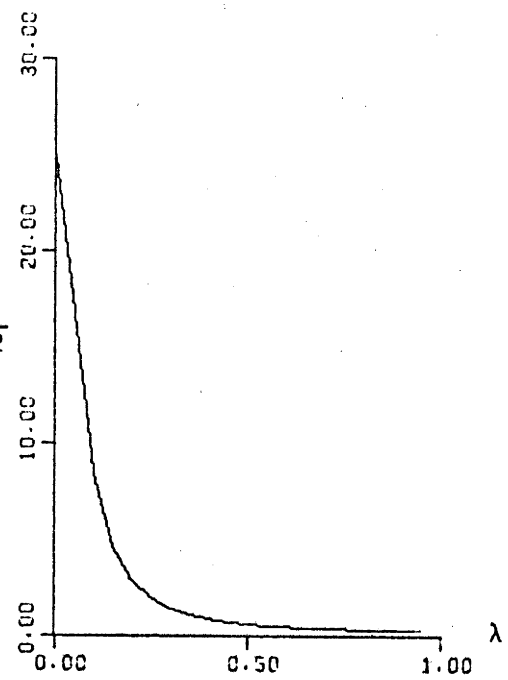
Table A8.2

Roots of A(L) = B(L)= = 0 for Models A-U

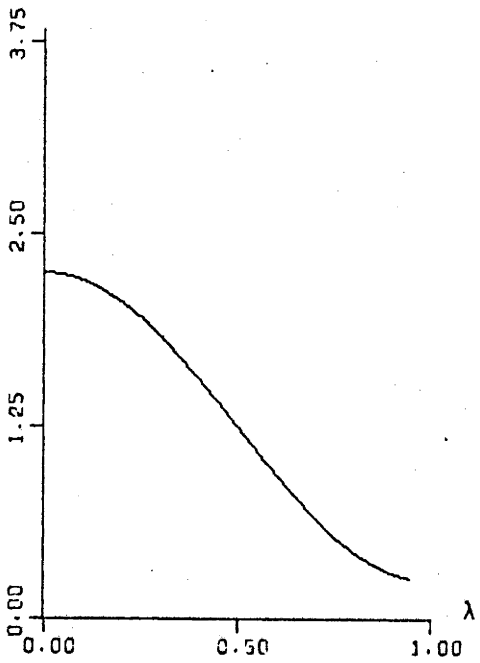
Model	B(L)	A(L)
A	1.2500	-2.0000
B	1.0833 + 0.7022i 1.0833 - 0.7022i	-1.0 + 1.5275i -1.0 - 1.5275i
C	1.4930 2.2108 + 2.3335i 2.2108 - 2.3335i	-3.4084 -0.5560 + 1.6788i -0.5560 - 1.6788i
D	1.0538 1.1280 + 1.7528i 1.1280 - 1.7528i	-1.3333 + 1.2472i -1.3333 - 1.2472i
E	1.2500 3.8333	-1.3333 + 1.2472i -1.3333 - 1.2472i
F	1.0538 1.1280 + 1.7528i 1.1280 - 1.7528i	-1.3333 + 1.2472i -1.3333 - 1.2472i
G	1.5129 + 0.2185i 1.5129 - 0.2185i	-2.0000
H	-5.0000 1.0833 + 0.7022i 1.0833 - 0.7022i	-10.0147 -1.0073 + 0.1525i -1.0073 - 0.1525i
I	-5.0000 1.0833 + 0.7022i 1.0833 - 0.7022i	-1.3333 + 1.2472i -1.3333 - 1.2472i
J	1.0538 1.1280 + 1.7528i 1.1280 - 1.7528i	-0.6667 + 1.1055i -0.6667 - 1.1055i
K	1.2500 3.8333	-1.3333 + 1.2472i -1.3333 - 1.2472i
L	1.5129 + 0.2185i 1.5129 - 0.2185i	1.5129 + 0.2185i 1.5129 - 0.2185i
M	1.2500 3.8333	1.2500 3.8333
N	1.5129 + 0.2185i 1.5129 - 0.2185i	1.3512 1.3512 1.3512 1.3512
G/S	1.5129 + 0.2185i 1.5129 - 0.2185i	-2.0000
O	1.6667 1.6667 1.6667	-3.3333 1.0933 1.0933 1.0933 1.0933
P	1.6667 1.6667 1.6667	-2.5000
G/NS	1.5129 + 0.2185i 1.5129 - 0.2185i	-2.0000
F/NS	1.0538 1.1280 + 1.7528i 1.1280 - 1.7528i	-1.3333 + 1.2472i -1.3333 - 1.2472i
Q	1.2500	-2.0000
R	1.0538 1.1280 + 1.7528i 1.1280 - 1.7528i	-2.0000
S	1.5129 + 0.2185i 1.5129 - 0.2185i	1.4286
T		-5.0000 1.2574 1.2574 1.2574 1.2574
U		-2.0000

SPECTRAL CHARACTERISTICS OF MODEL A

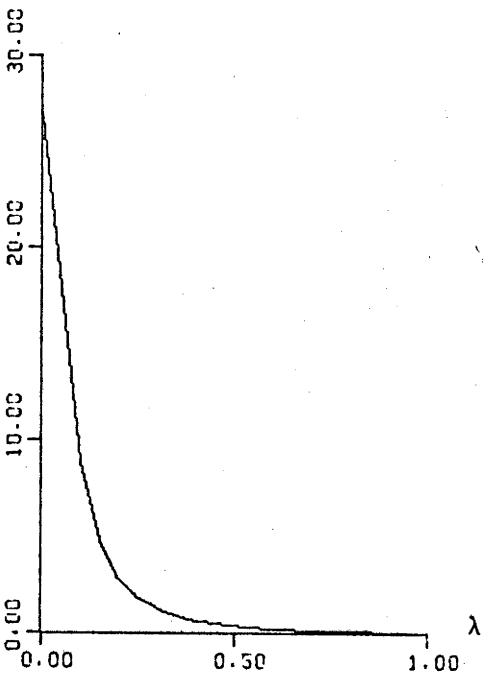
$\frac{1}{|B(\lambda)|^2}$



$|A(\lambda)|^2$

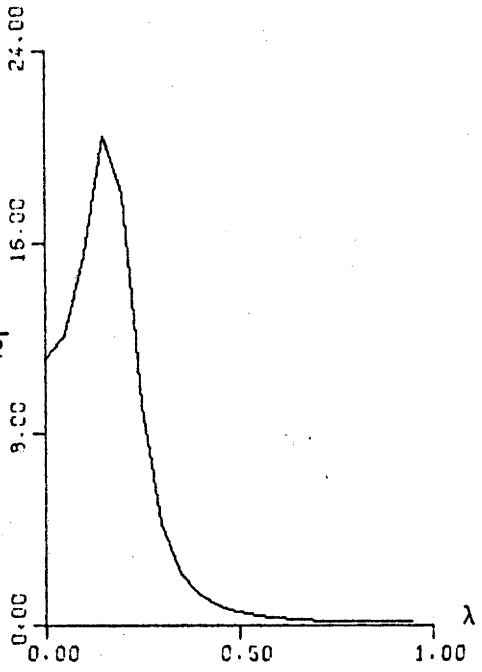


$f_{yy}(\lambda)$

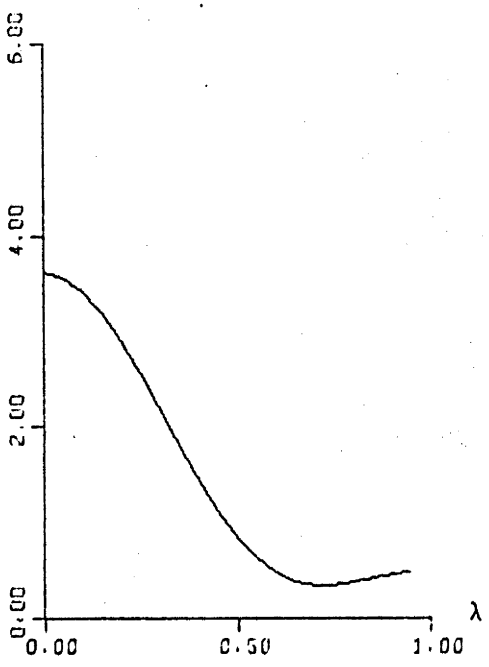


SPECTRAL CHARACTERISTICS OF MODEL B

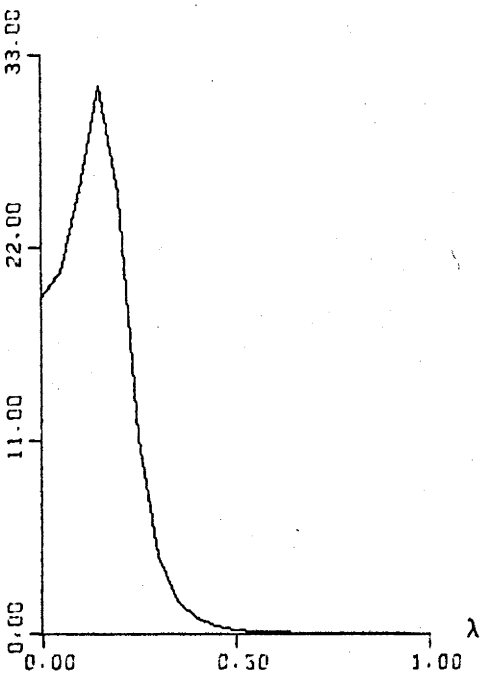
$\frac{1}{|B(\lambda)|^2}$



$|A(\lambda)|^2$

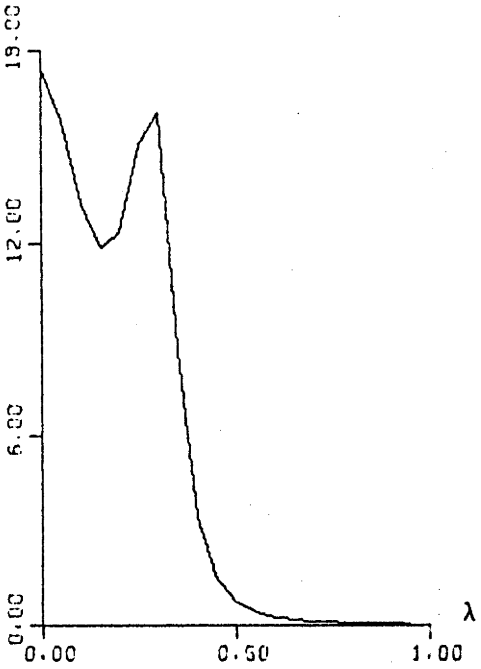


$f_{yy}(\lambda)$

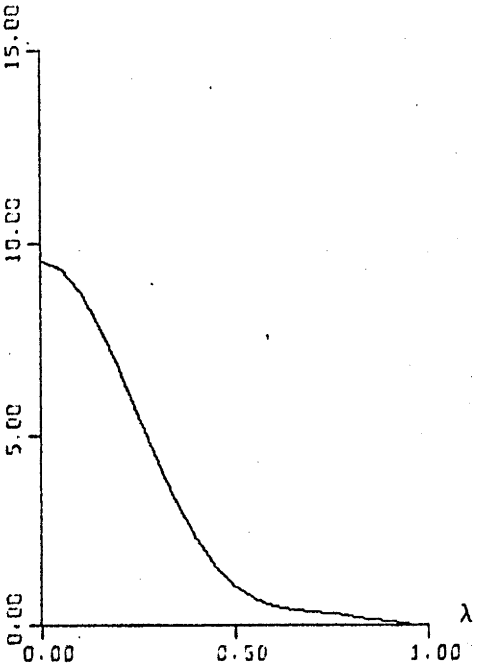


SPECTRAL CHARACTERISTICS OF MODEL C

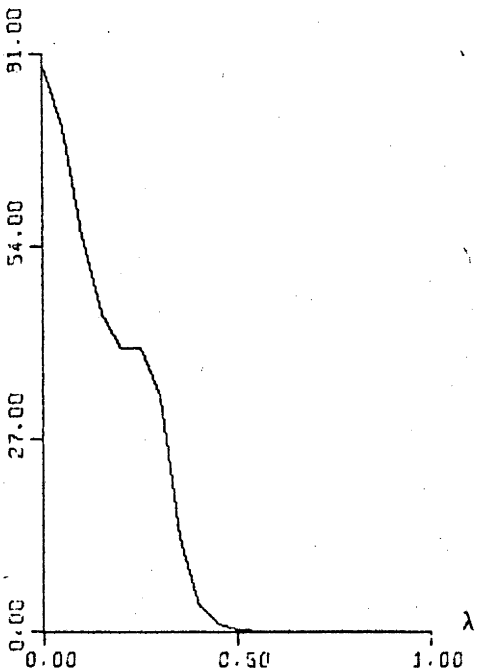
$\frac{1}{|B(\lambda)|^2}$



$|A(\lambda)|^2$

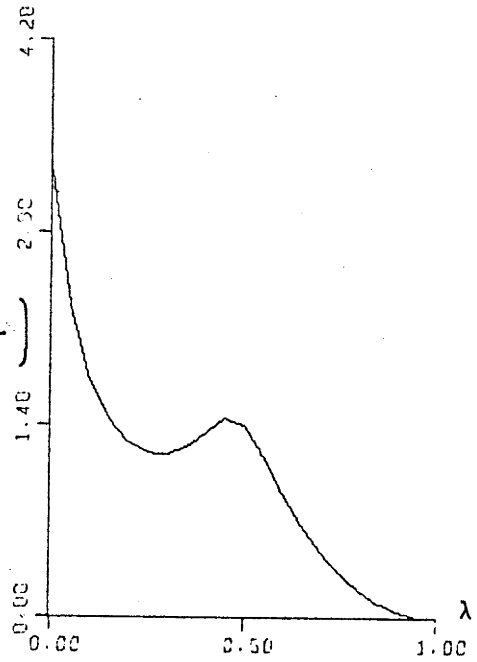


$f_{yy}(\lambda)$

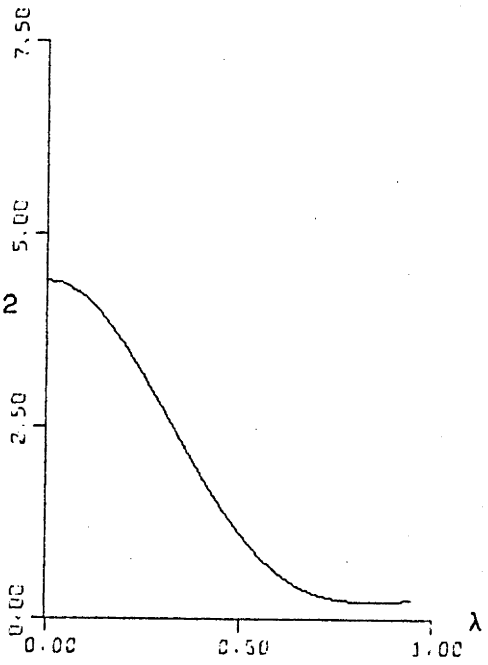


SPECTRAL CHARACTERISTICS OF MODEL D

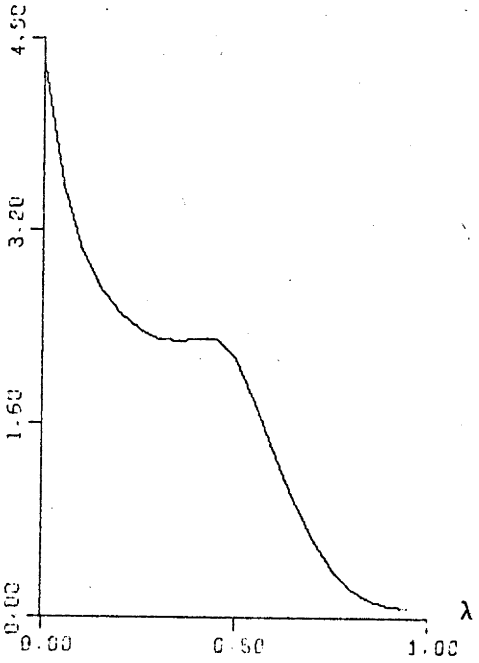
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



$|A(\lambda)|^2$

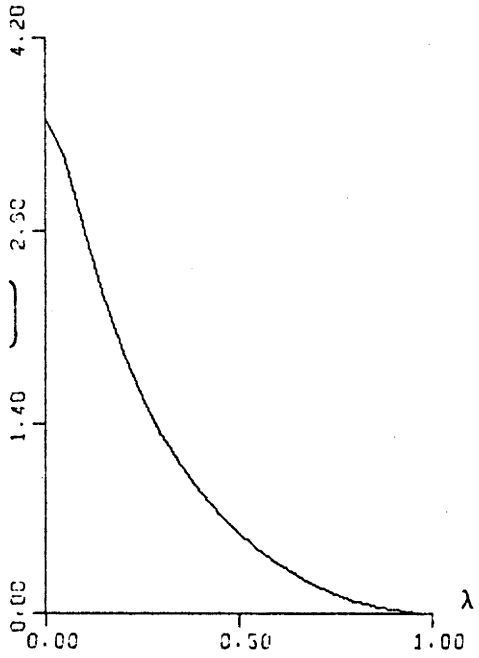


$\log(f_{yy}(\lambda))$

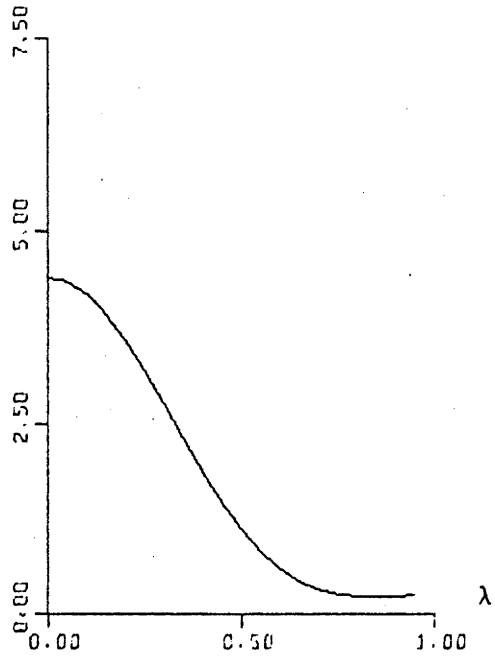


SPECTRAL CHARACTERISTICS OF MODEL E

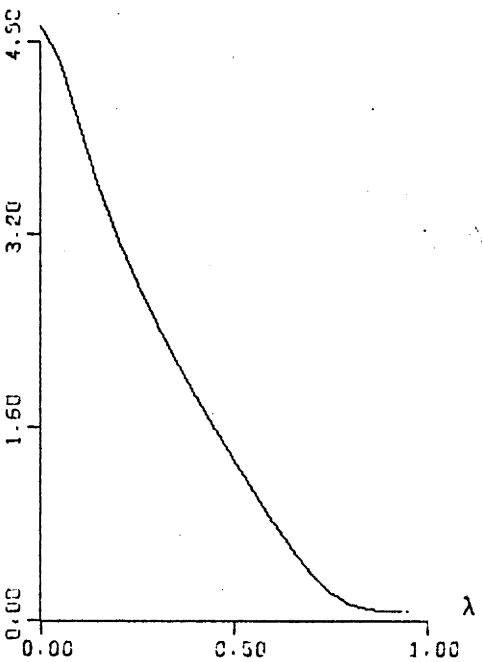
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



$|A(\lambda)|^2$

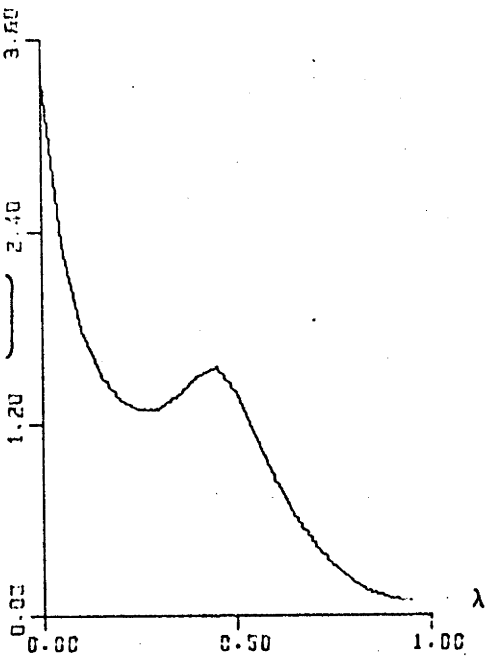


$\log(f_{yy}(\lambda))$

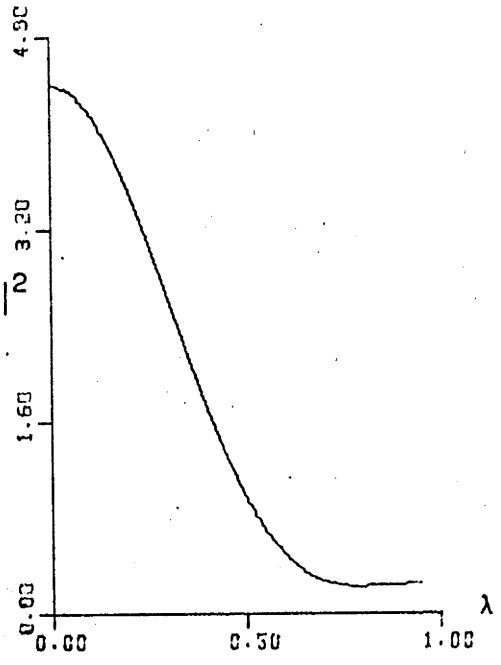


SPECTRAL CHARACTERISTICS OF MODEL F

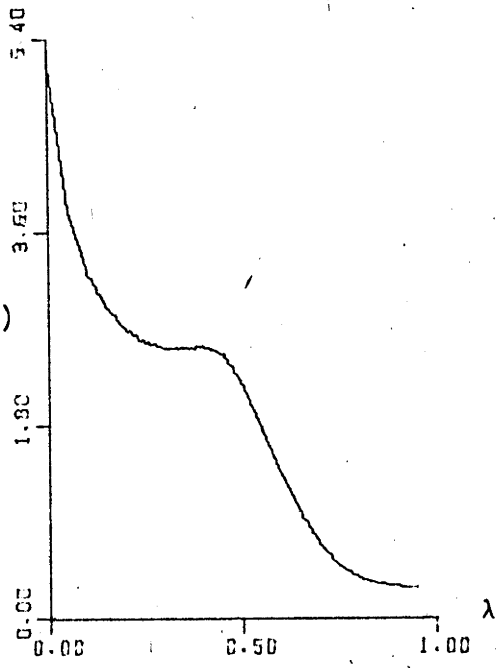
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



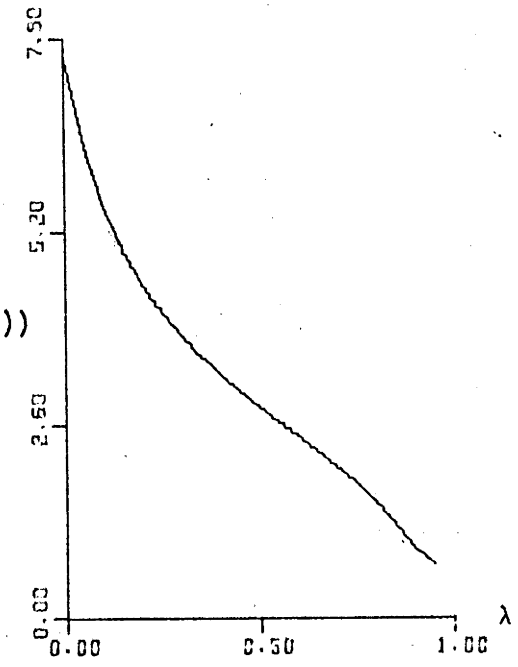
$|A(\lambda)|^2$



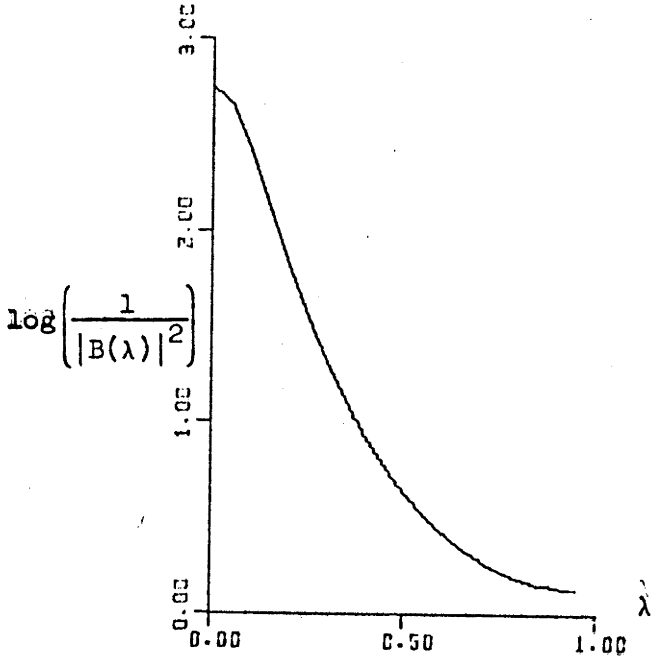
$\log(f_{yy}(\lambda))$

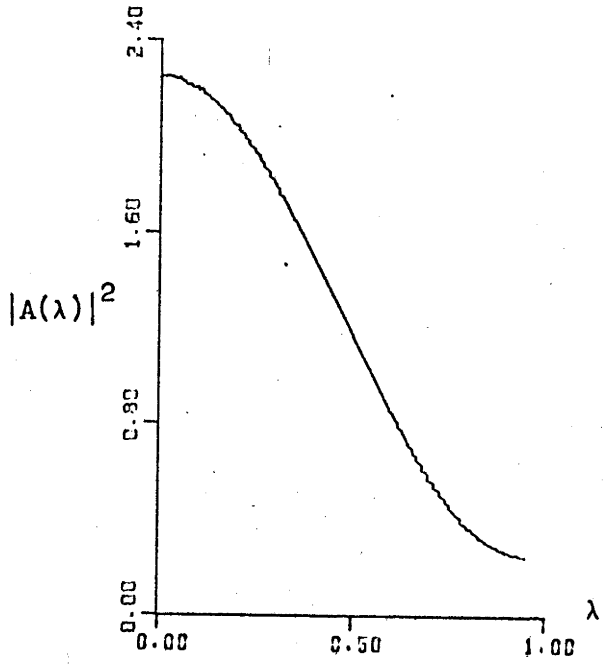


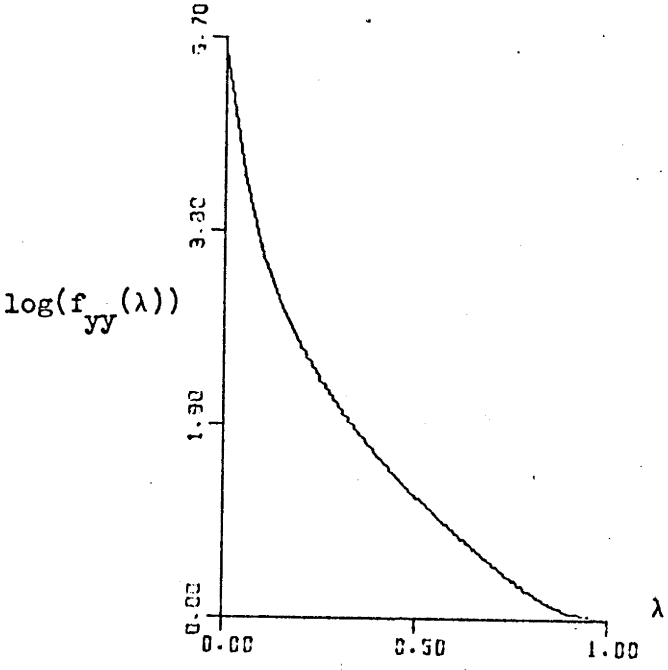
$\log(f_{xx}(\lambda))$

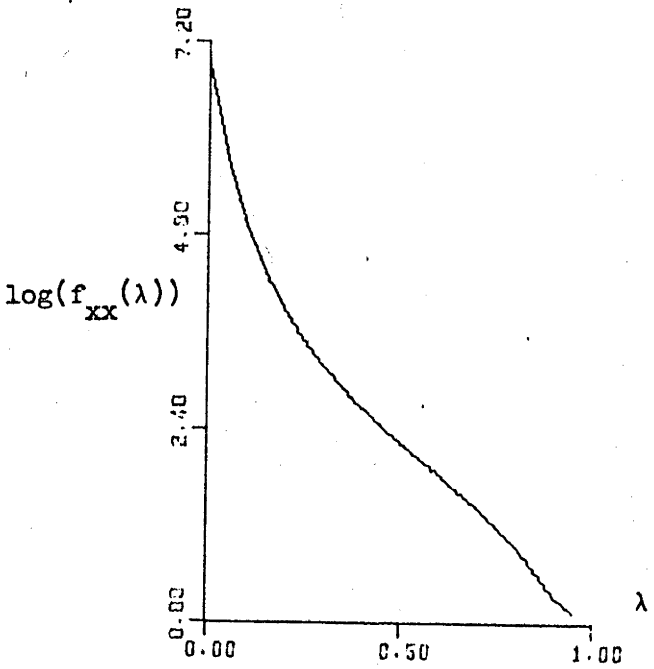


SPECTRAL CHARACTERISTICS OF MODEL G



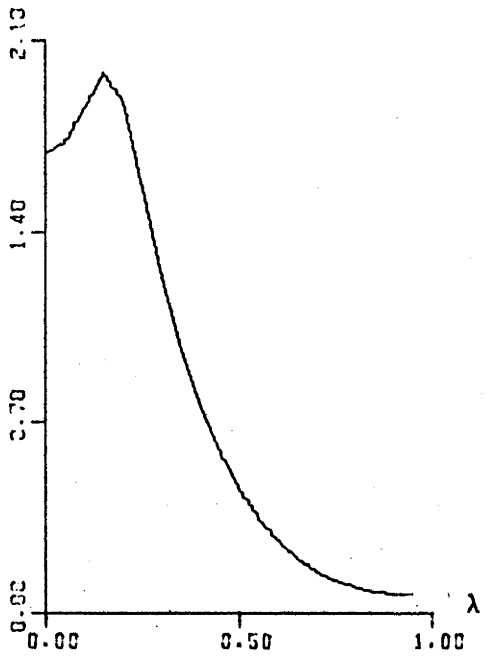




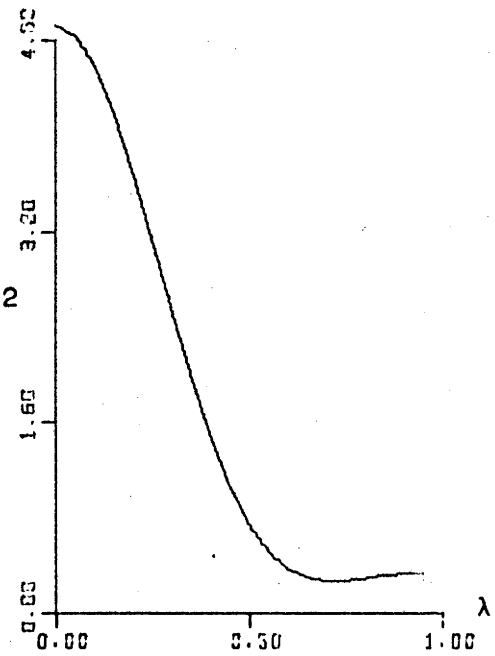


SPECTRAL CHARACTERISTICS OF MODEL H

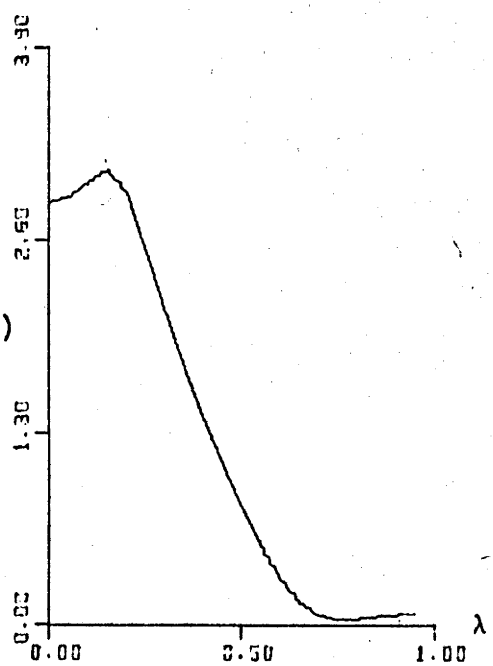
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



$|A(\lambda)|^2$

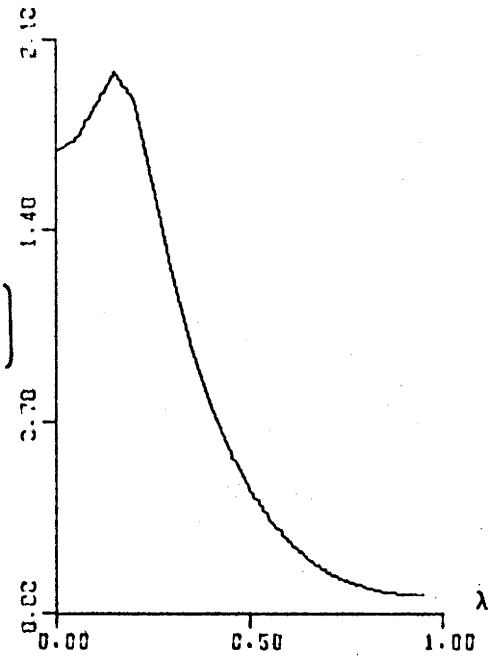


$\log(f_{yy}(\lambda))$

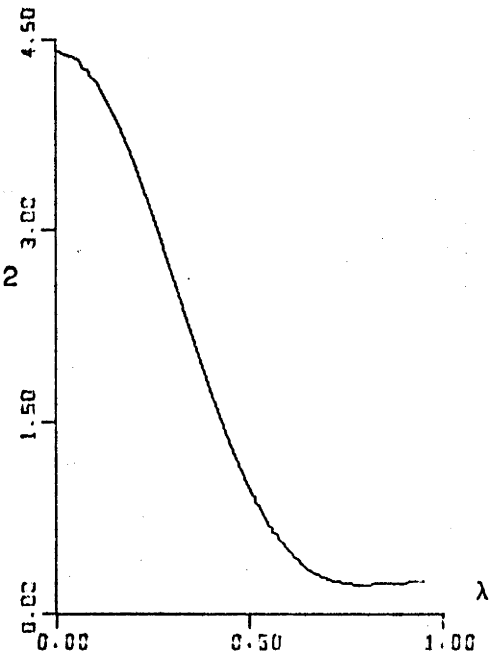


SPECTRAL CHARACTERISTICS OF MODEL I

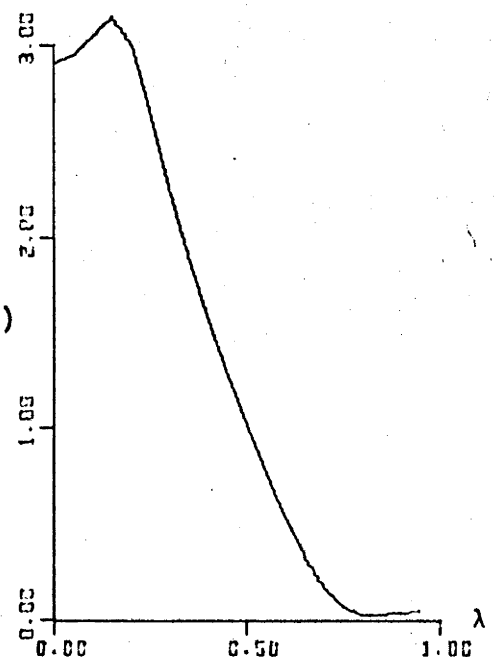
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



$|A(\lambda)|^2$

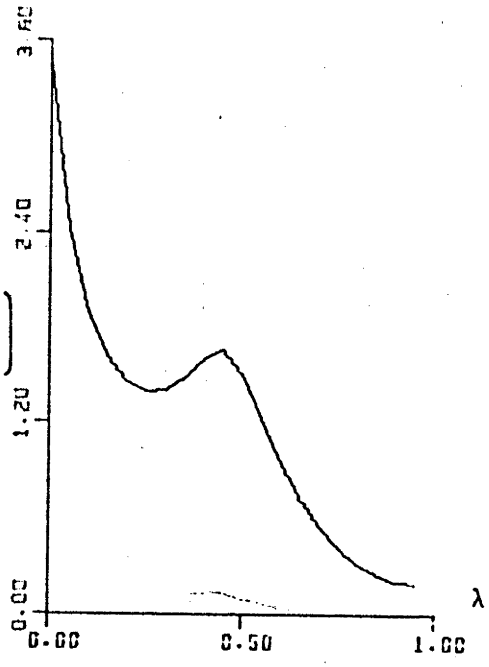


$\log(f_{yy}(\lambda))$

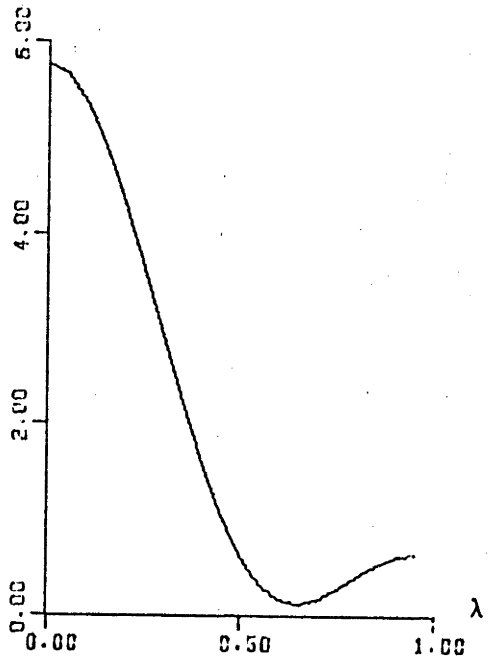


SPECTRAL CHARACTERISTICS OF MODEL J

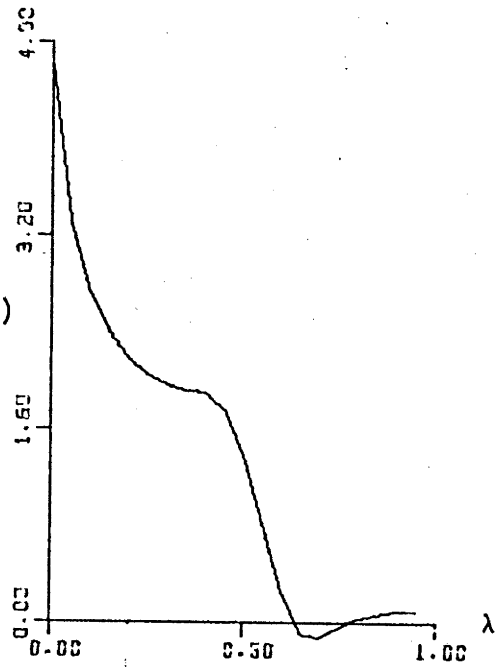
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



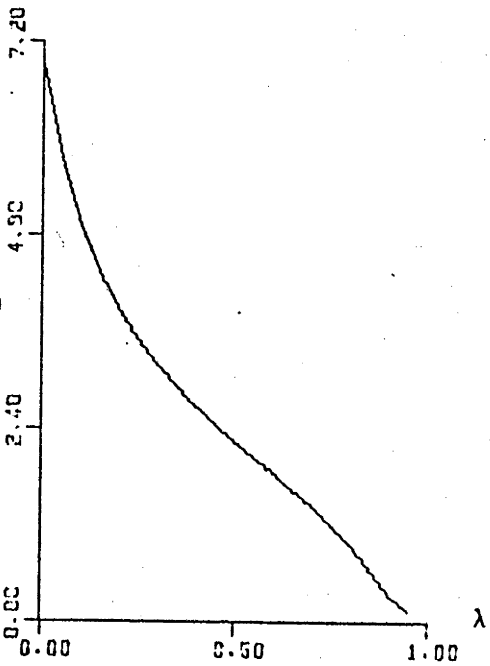
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

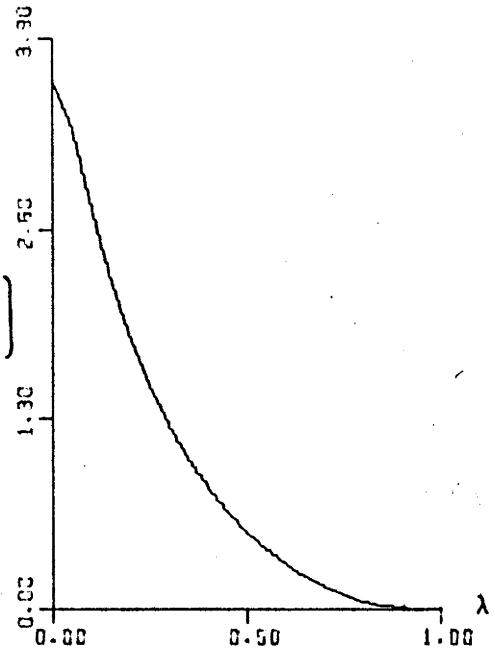


$\log(f_{xx}(\lambda))$

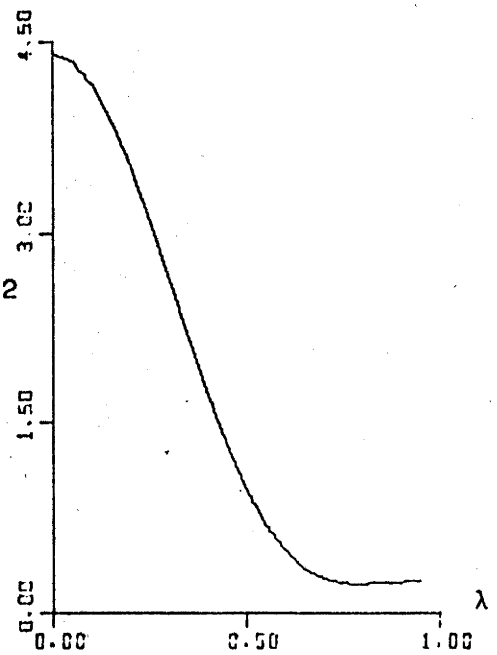


SPECTRAL CHARACTERISTICS OF MODEL K

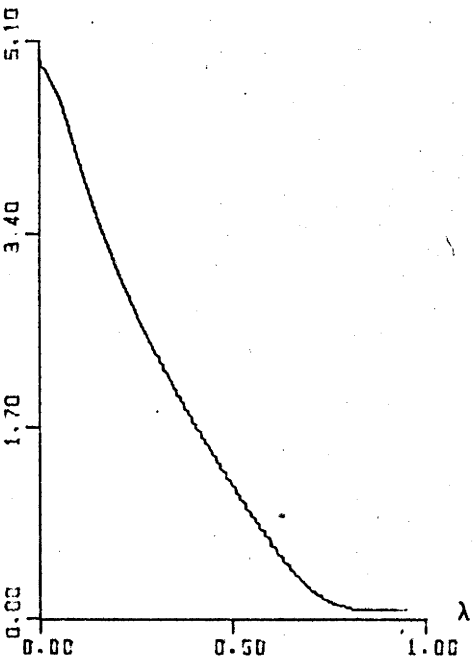
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



$|A(\lambda)|^2$

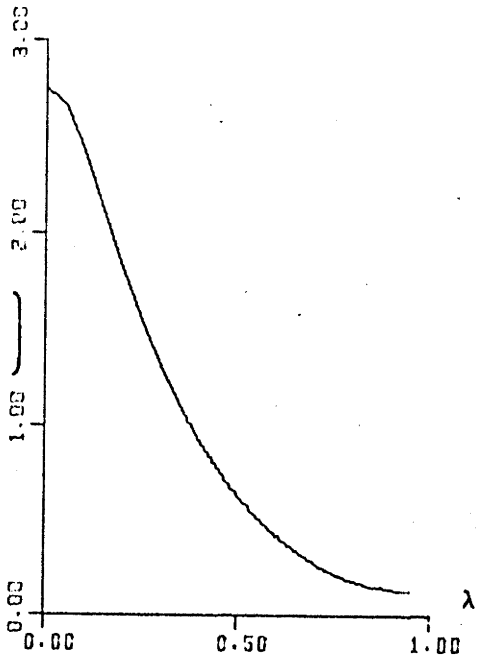


$\log(f_{yy}(\lambda))$

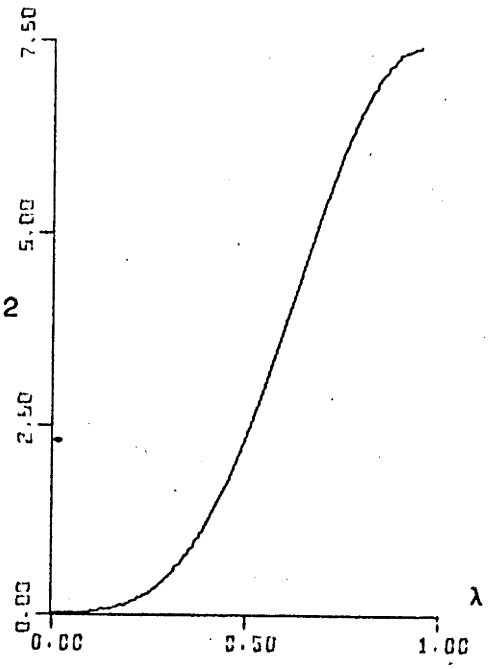


SPECTRAL CHARACTERISTICS OF MODEL L

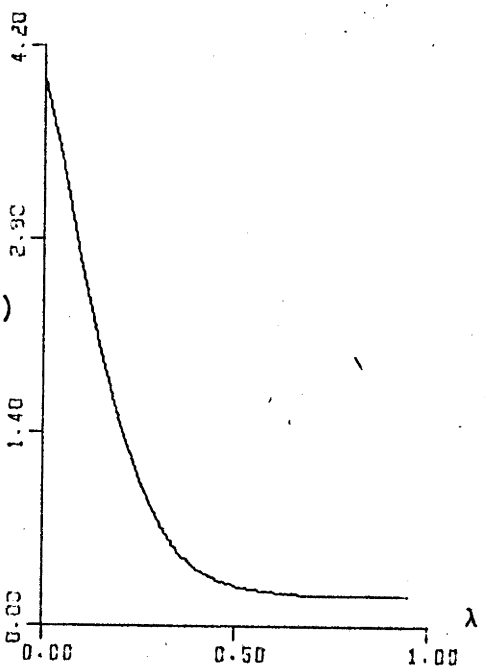
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



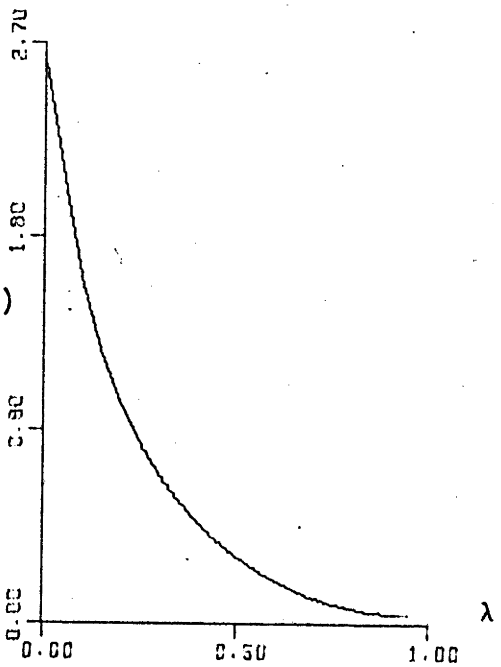
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

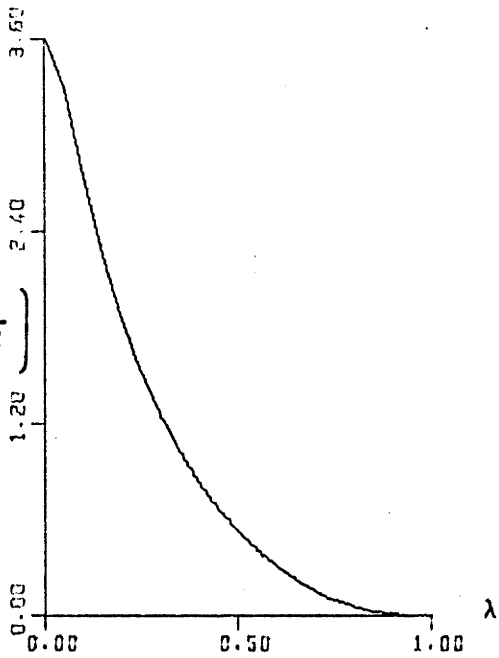


$\log(f_{xx}(\lambda))$

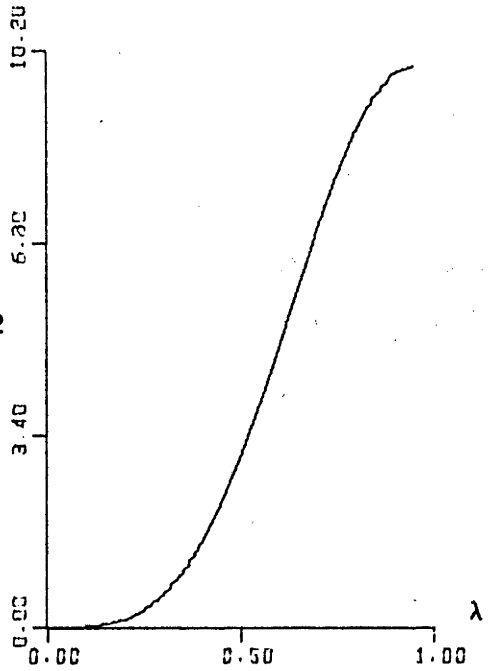


SPECTRAL CHARACTERISTICS OF MODEL M

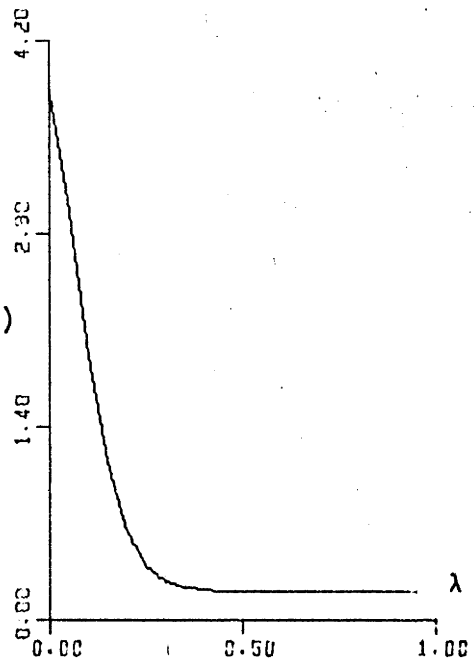
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



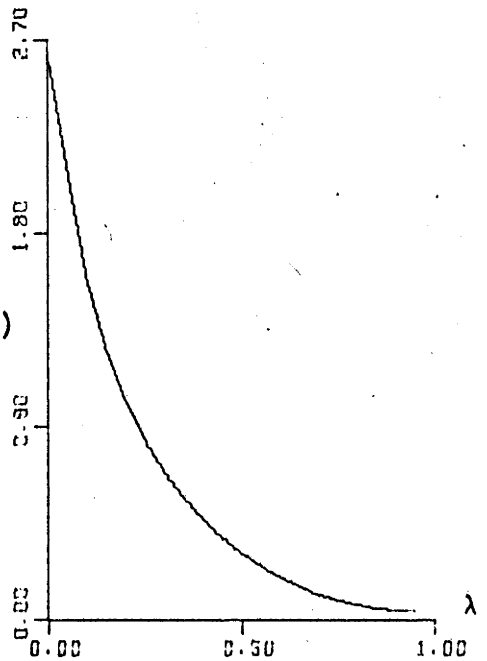
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

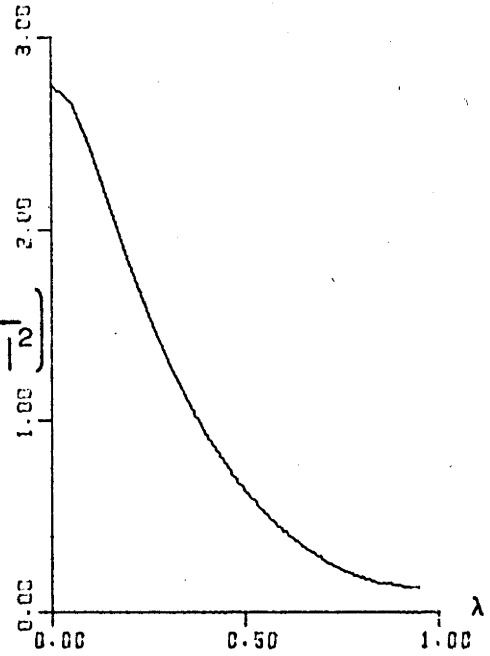


$\log(f_{xx}(\lambda))$

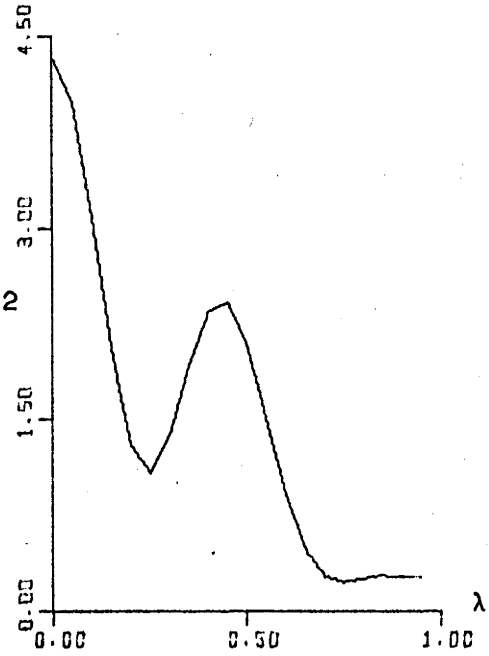


SPECTRAL CHARACTERISTICS OF MODEL N

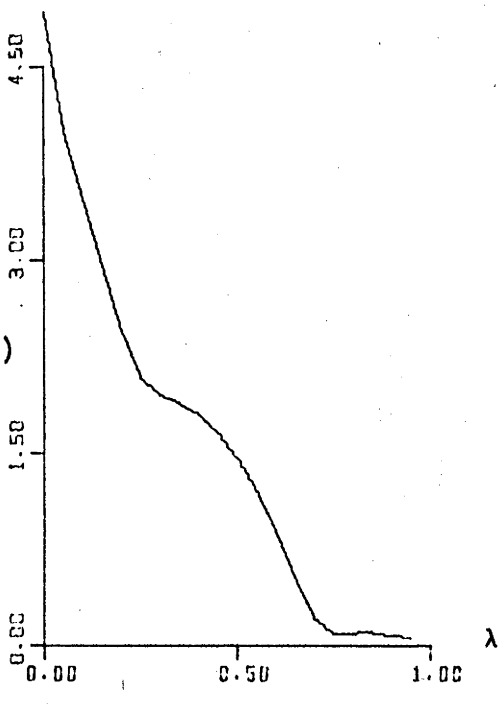
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



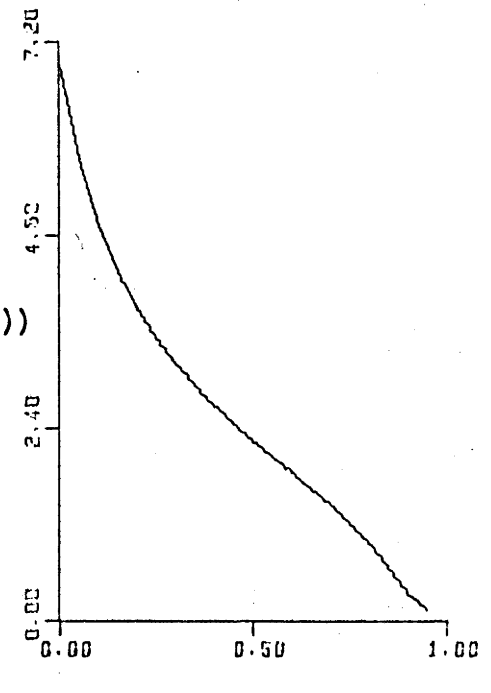
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

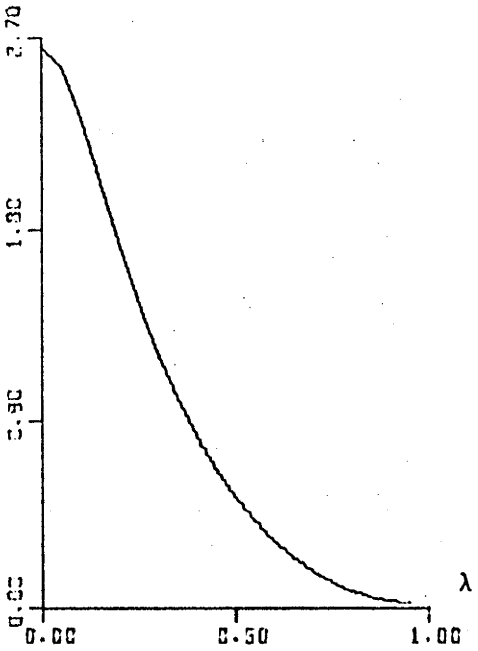


$\log(f_{xx}(\lambda))$

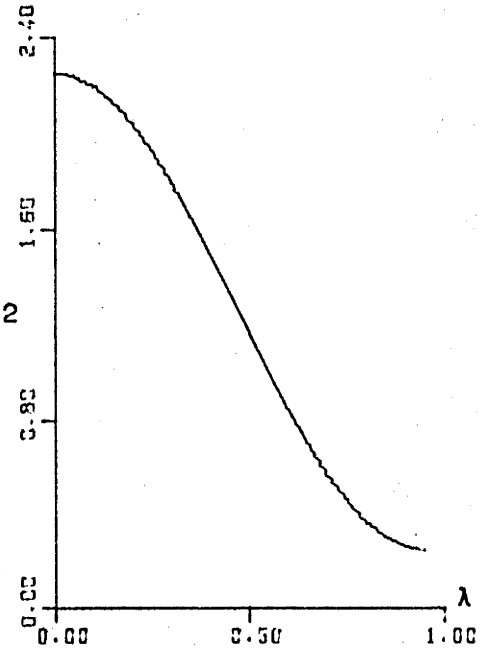


SPECTRAL CHARACTERISTICS OF MODEL G/S

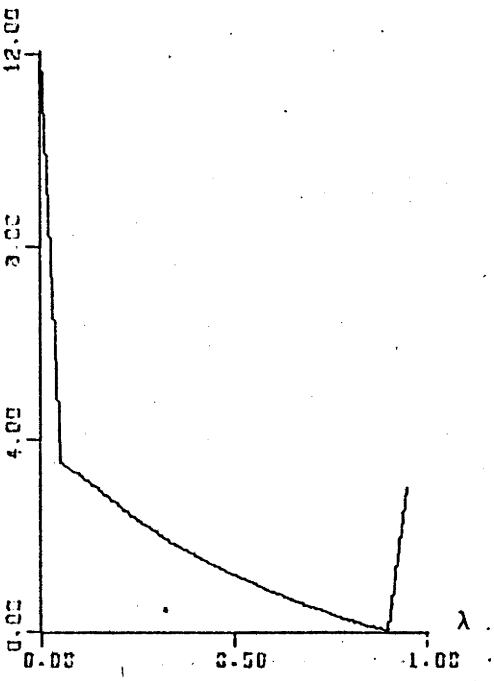
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



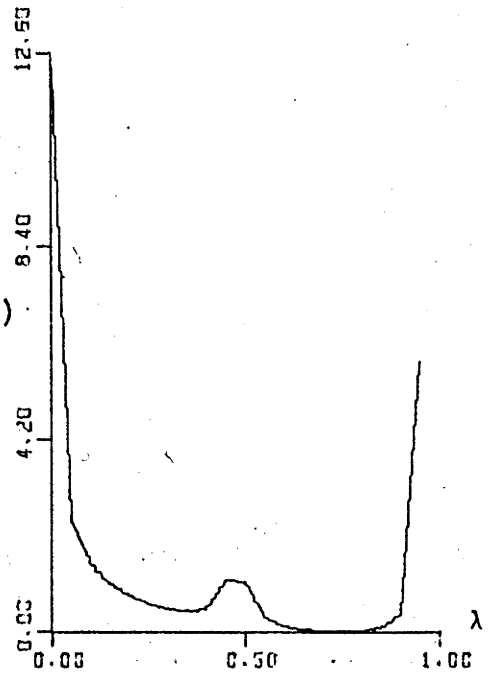
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

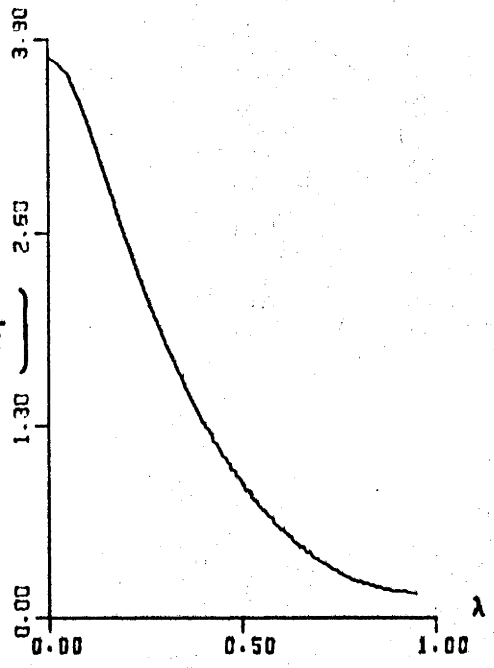


$\log(f_{xx}(\lambda))$

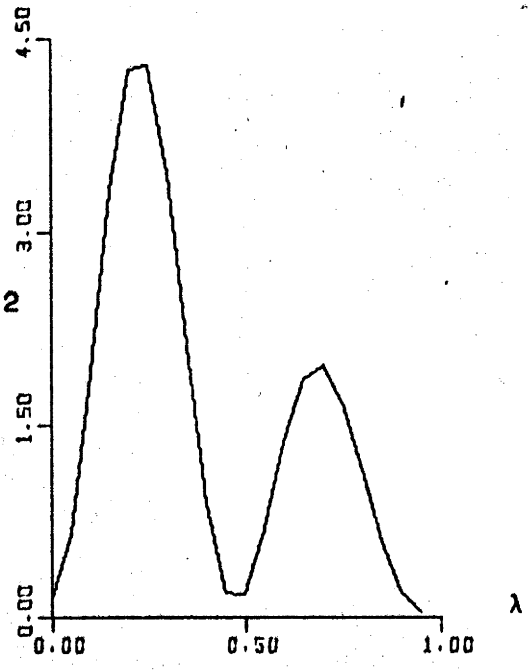


SPECTRAL CHARACTERISTICS OF MODEL 0

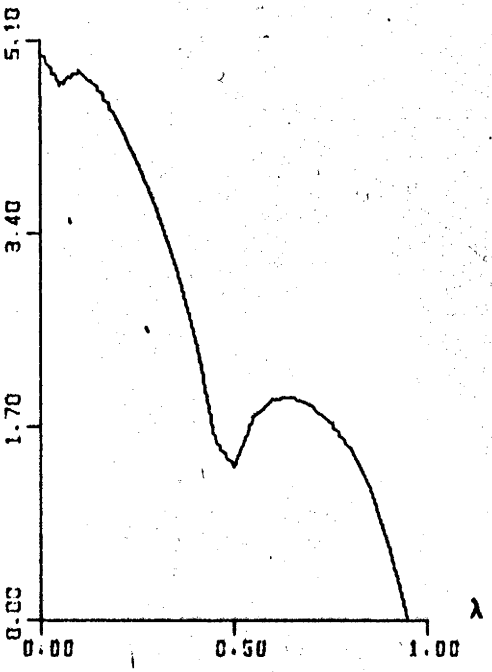
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



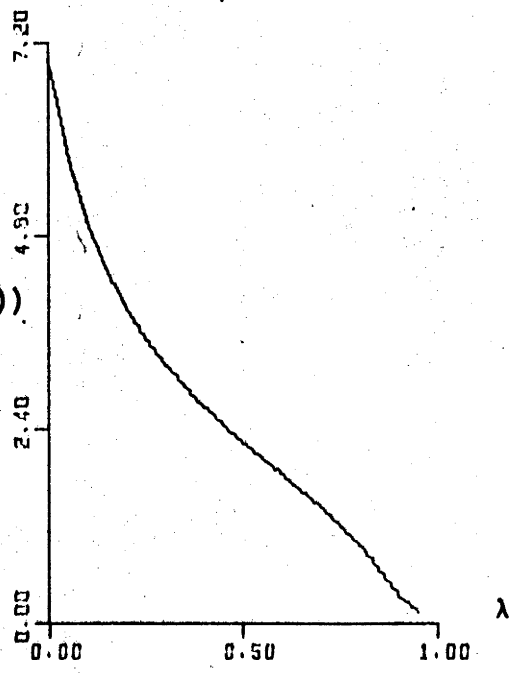
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

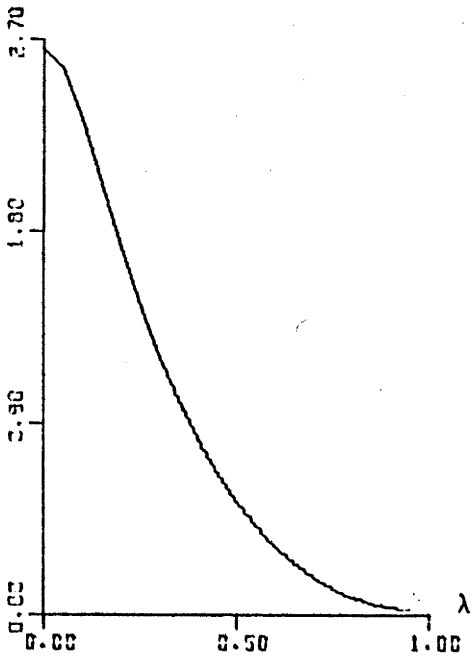


$\log(f_{xx}(\lambda))$

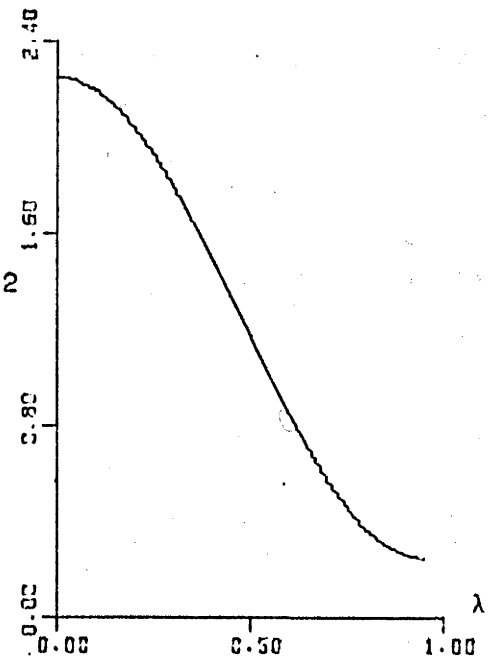


SPECTRAL CHARACTERISTICS OF MODEL P

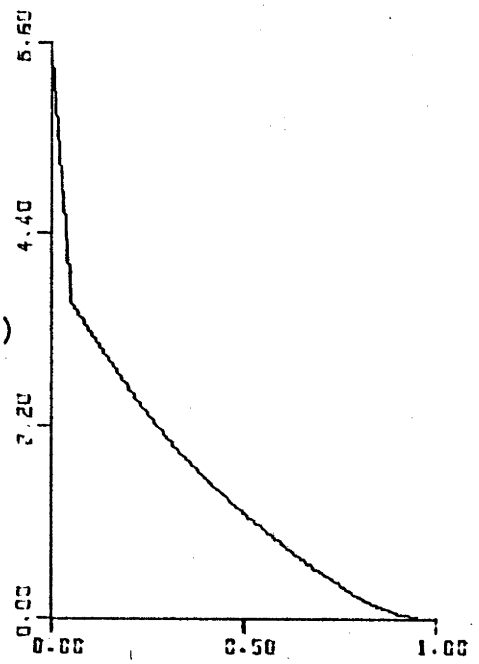
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



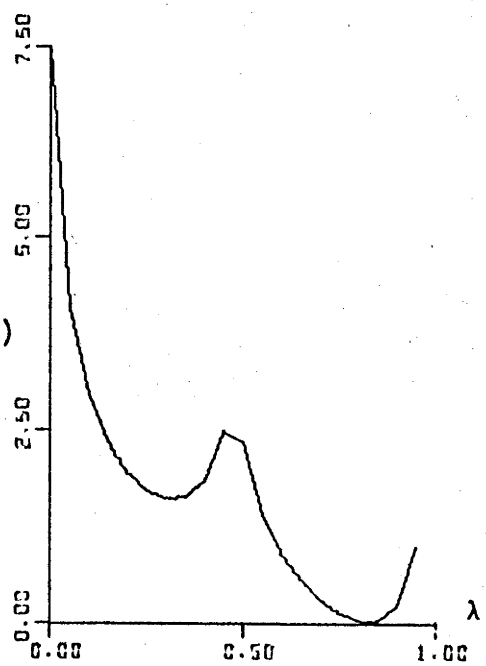
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

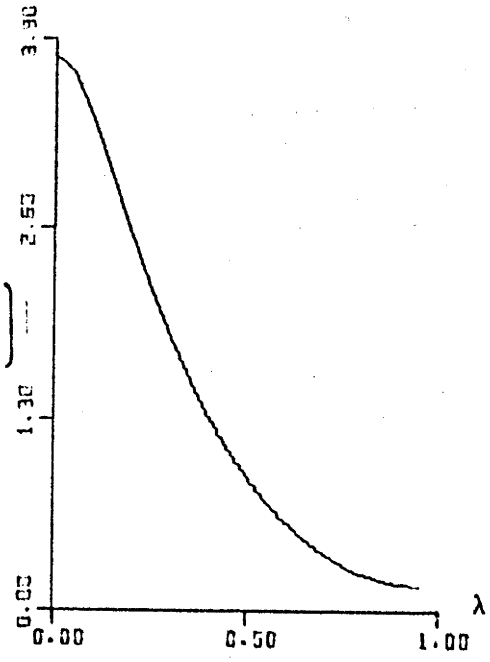


$\log(f_{xx}(\lambda))$

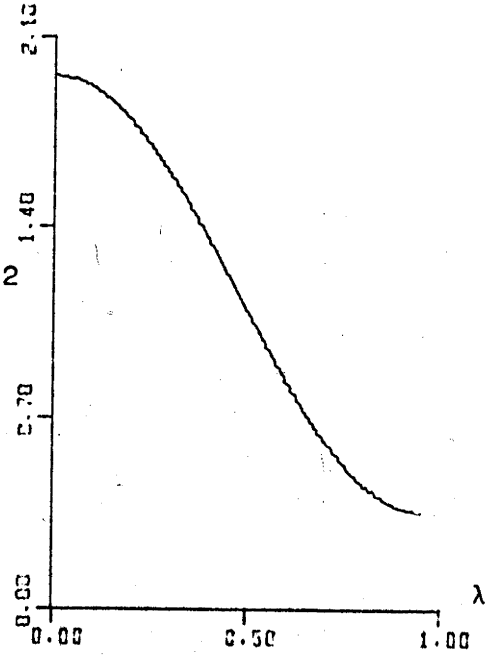


SPECTRAL CHARACTERISTICS OF MODEL G/NS

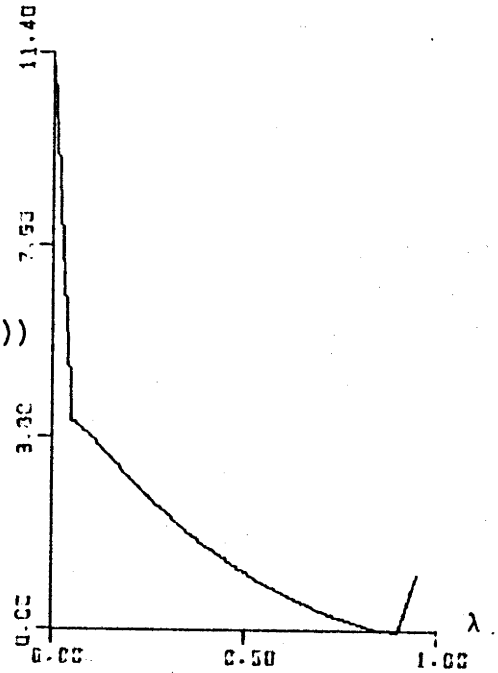
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



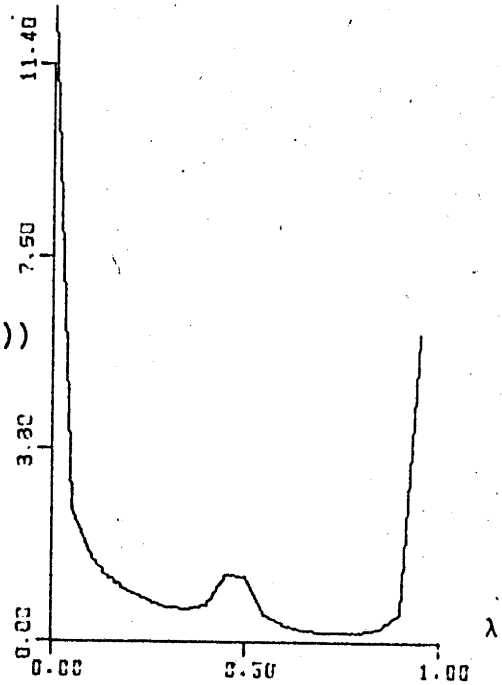
$|A(\lambda)|^2$



$\log\,(f_{yy}(\lambda))$

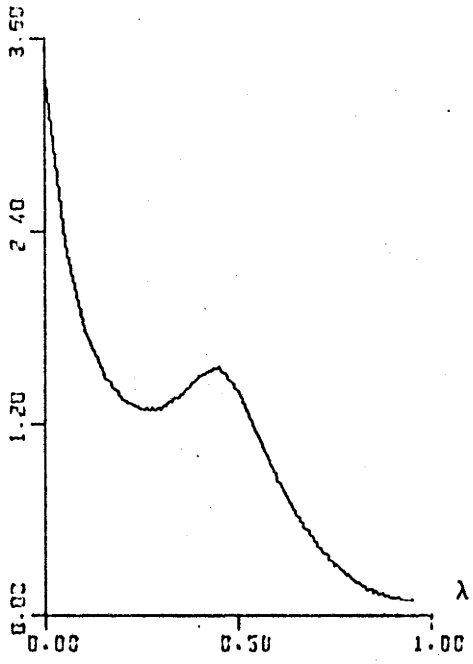


$\log(f_{xx}(\lambda))$

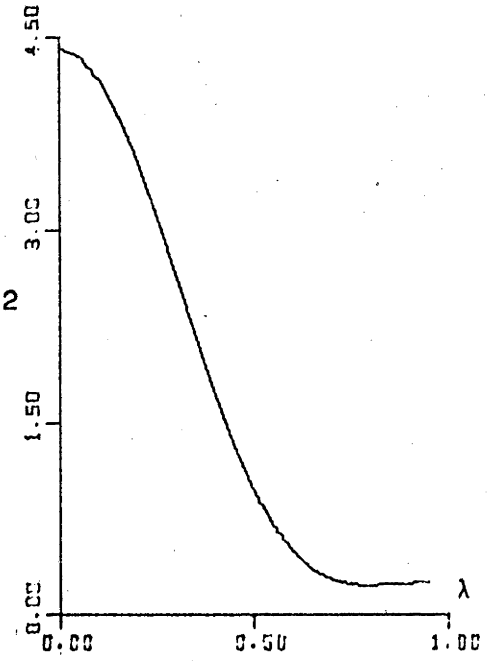


SPECTRAL CHARACTERISTICS OF MODEL F/NS

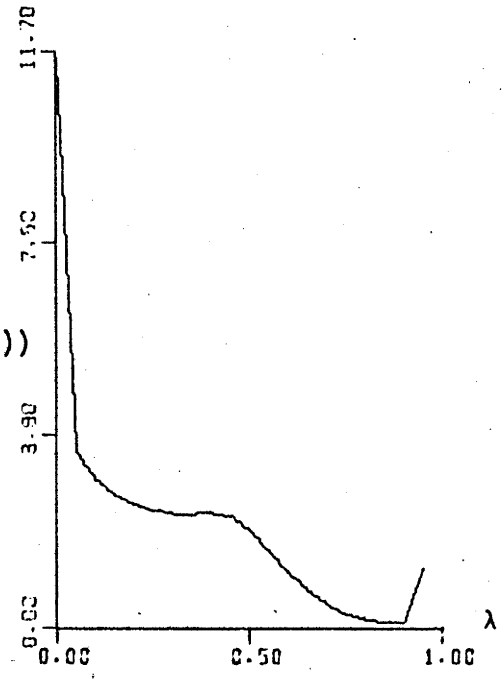
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



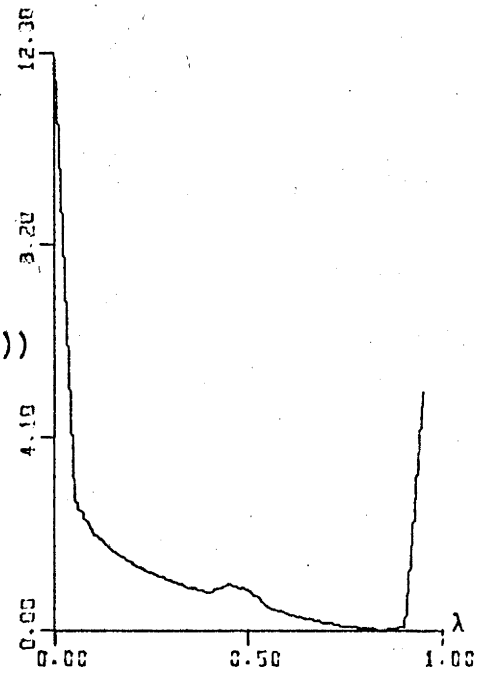
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

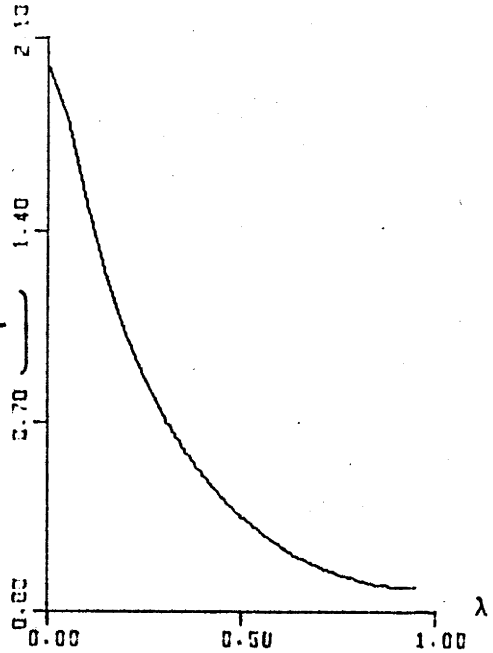


$\log(f_{xx}(\lambda))$

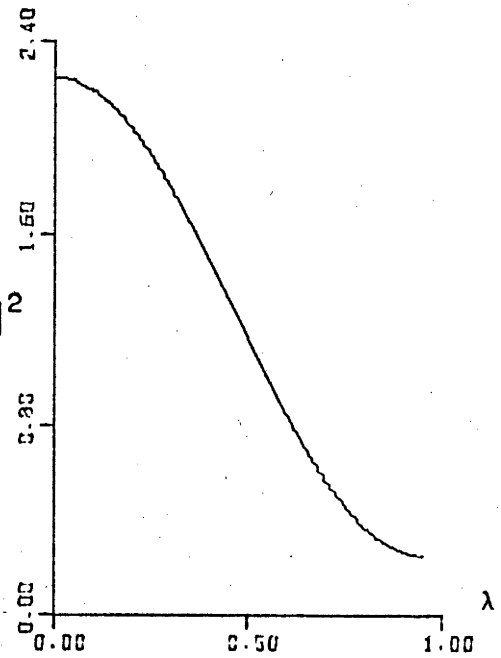


SPECTRAL CHARACTERISTICS OF MODEL Q

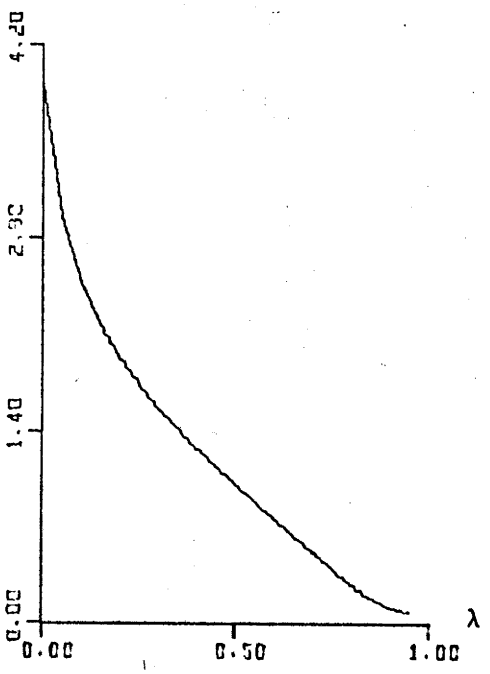
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



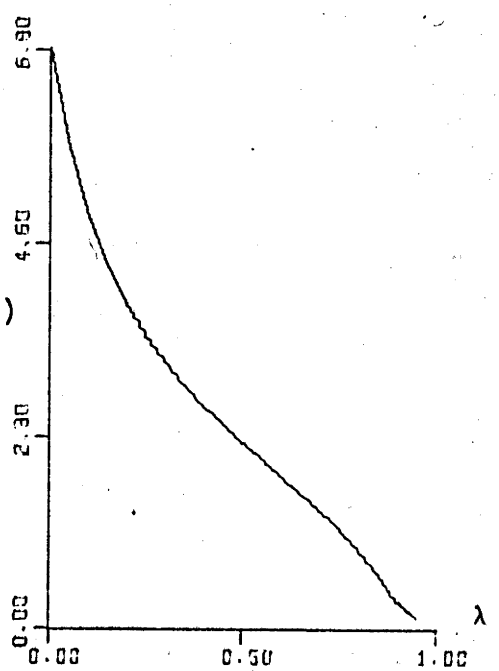
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

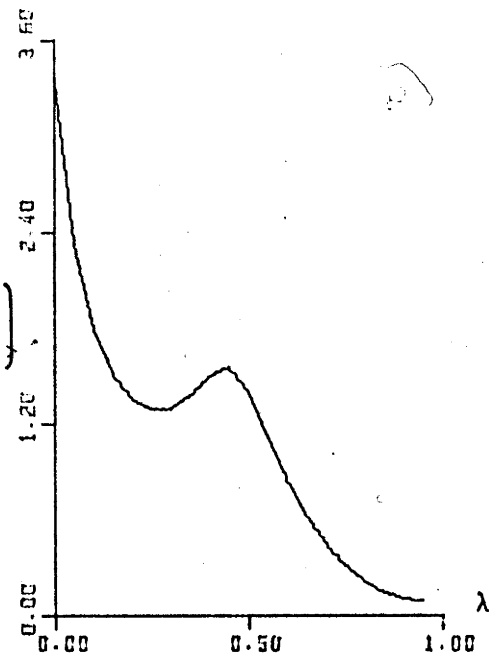


$\log(f_{xx}(\lambda))$

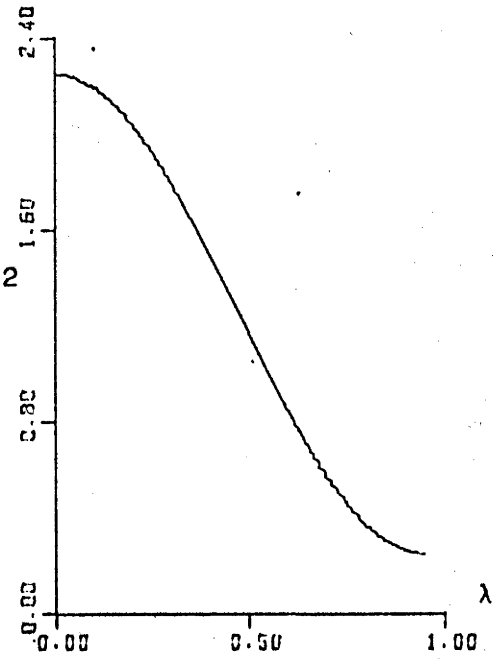


SPECTRAL CHARACTERISTICS OF MODEL R

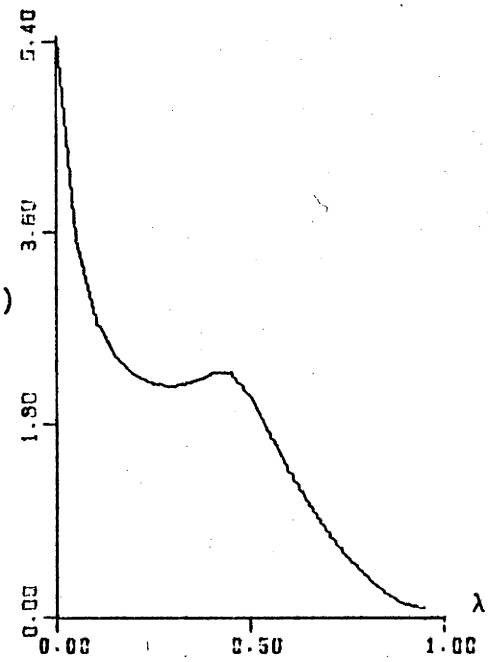
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



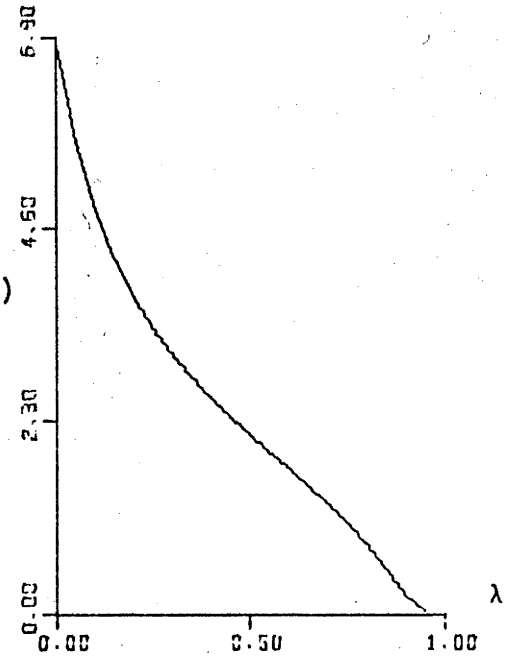
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

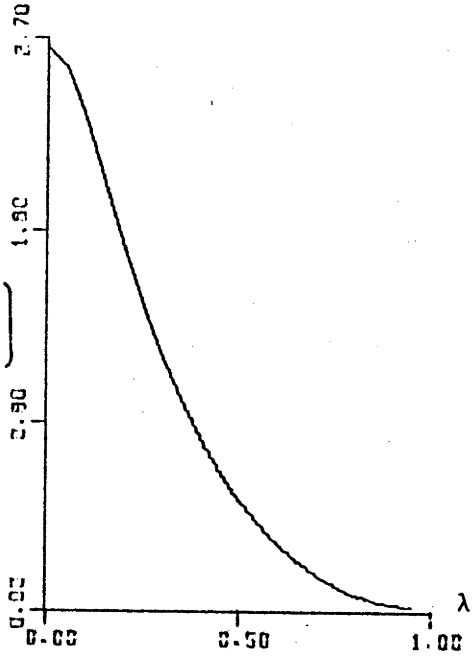


$\log(f_{xx}(\lambda))$

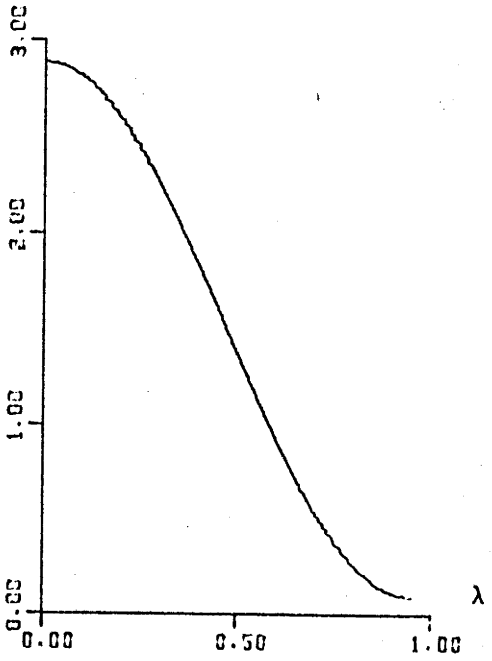


SPECTRAL CHARACTERISTICS OF MODEL S

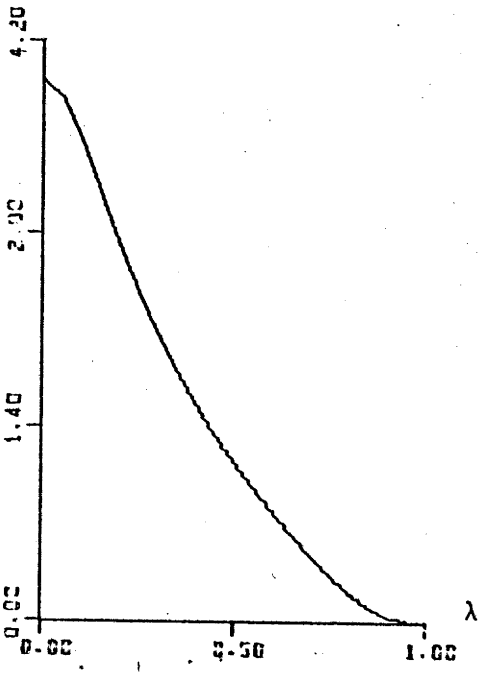
$\log\left(\frac{1}{|B(\lambda)|^2}\right)$



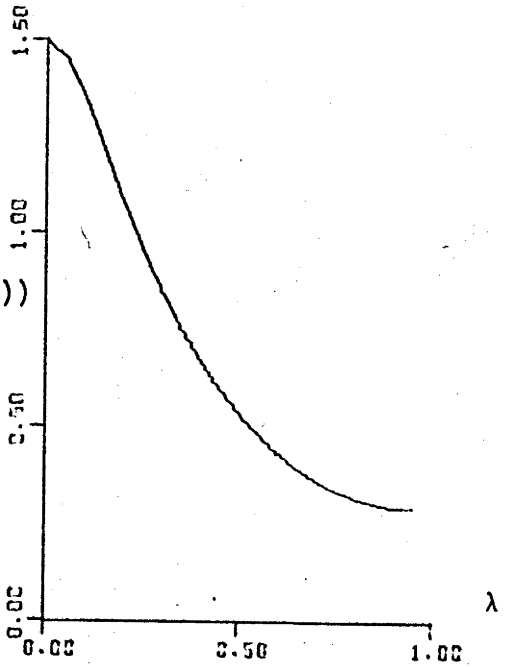
$|A(\lambda)|^2$



$\log(f_{yy}(\lambda))$

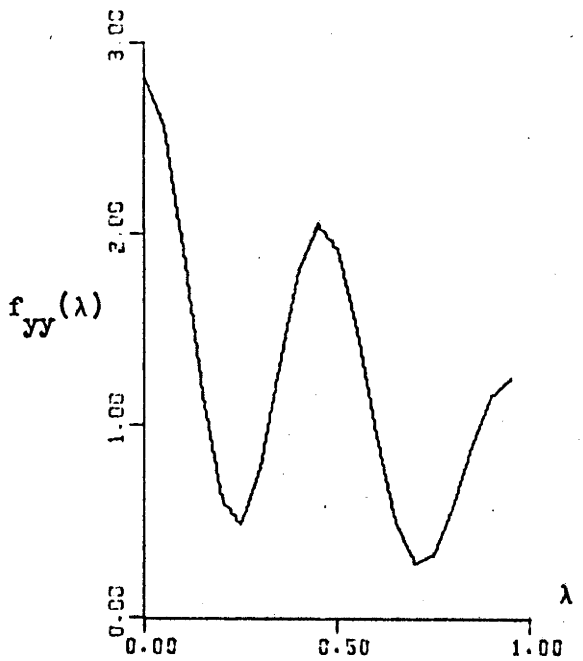


$\log(f_{xx}(\lambda))$

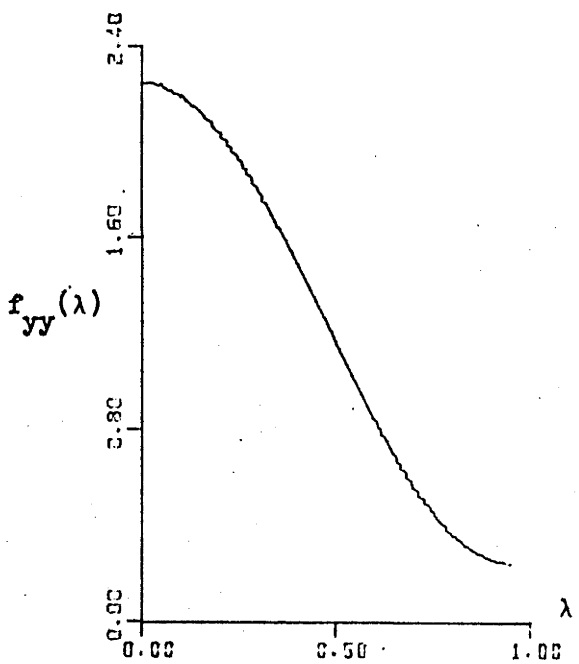


SPECTRAL CHARACTERISTICS OF MODELS T AND U

MODEL T

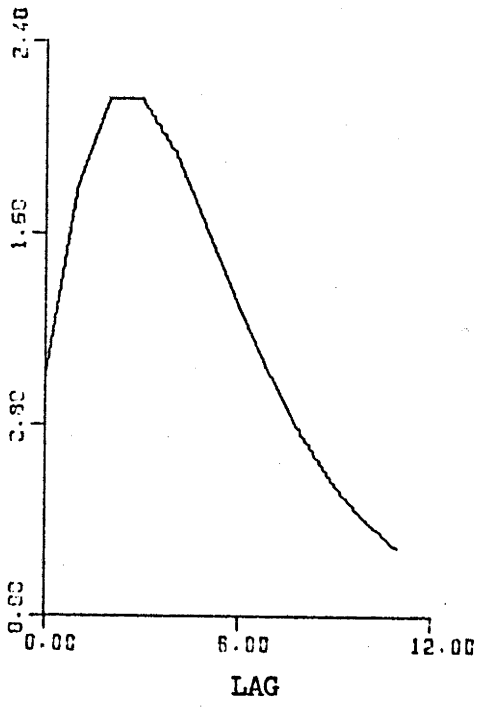


MODEL U

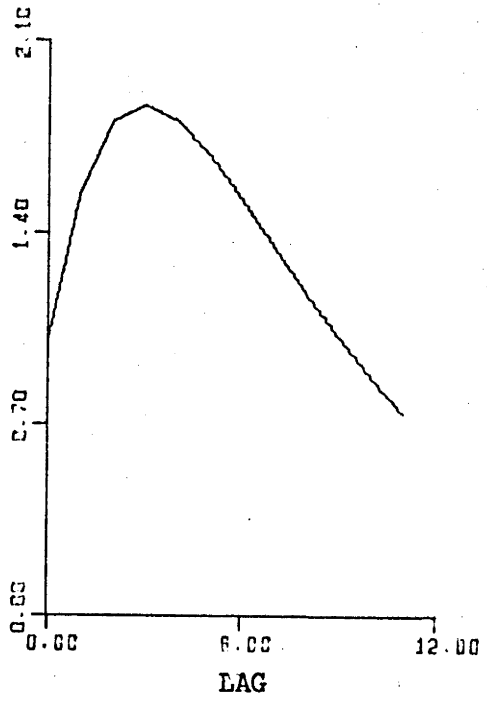


LAG DISTRIBUTIONS FOR VARIOUS MODELS

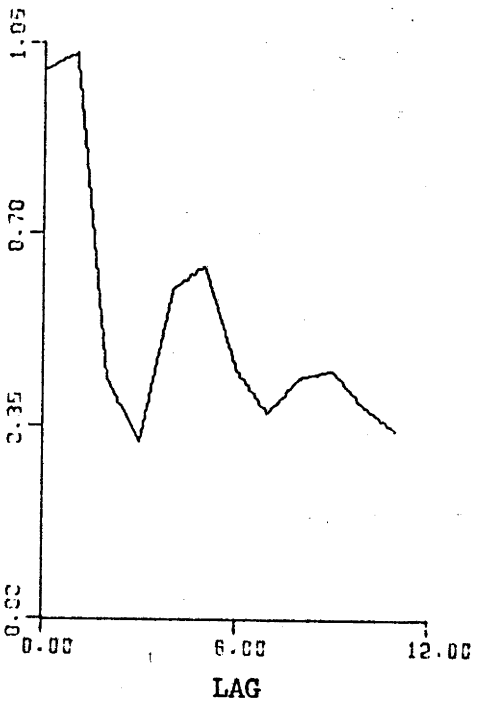
$B(L) = (1-0.6L)^3$



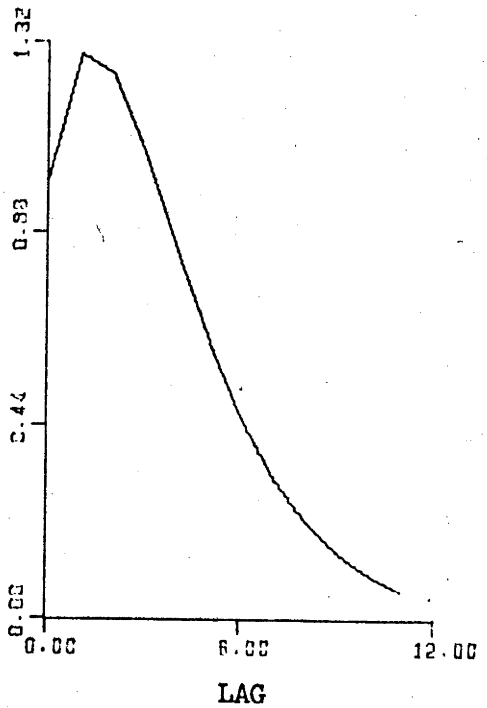
$B(L) = 1-1.55L + 0.6L^2$



$B(L) = 1-1.031L+0.630L^2-0.524L^3$



$B(L) = 1-1.295L+0.428L^2$



CHAPTER 4

A Comparison of Some Alternative Estimators of ARMAX Models

4.1 Introduction

Chapter 2 has described a number of estimators that are proposed in the econometric literature for the estimation of ARMAX models. One of these - the Phillips estimator - was subjected to a Monte Carlo investigation in Chapter 3 in an attempt to gain some appreciation of its small sample performance and the conditions under which it might be expected to perform poorly. With some reservations it seemed that if the estimator was applied in samples of around 70 asymptotic theorems would hold.

This chapter will evaluate alternative estimators by the same technique in order to achieve a ranking to indicate which should be used in empirical analysis. That this ranking may be a function of sample size can be seen by comparing the degrees of freedom remaining with each estimator so that small sample biases may vary between estimators. The six estimators to be examined are:-

(i) Phillips' with e* estimated (PH).

(ii) Phillips' with e* set at zero (PH/0).

(iii) Aitken Estimator in the Time Domain (A.T.D.).

(iv) Box-Jenkins' Estimator (B-J).

(v) Ordinary Least Squares (O.L.S.).

(vi) Aitken Estimator in the Frequency Domain (A.F.D.).

All comparisons in Sections 4.2 - 4.6 will be pairwise with the Phillips estimator (PH) as the Monte Carlo studies of Chapter 3 have highlighted its strengths and weaknesses and it can therefore serve in

the nature of a benchmark. Furthermore, any comparison will be not only of distributional properties but also computational time as the latter will be of considerable importance to econometric exercises in which a wide variety of specifications must be tested. Finally Section 4.7 analyses two economic time series and attempts to parameterize these by an application of the five estimators. One of the series considered exhibits some peculiarities that will be discussed in Chapter 6.

The development of programs for the six estimators was not a trivial task and it was quickly found that Monte Carlo experimentation with a variety of estimators is a very time consuming occupation (particularly with the multiple starting values used for each algorithm), so that the range and number of such experiments is quite limited. Nevertheless some attempt was made to select "realistic" models and in general it is felt that the comparative results would extend to other simulations. Where there is some doubt specific mention will be made of possible methods for testing this conclusion.

4.2 A comparison of A.T.D. and PH

Three ARMAX models were selected to compare A.T.D. and PH, the main source of variation being the shape of the squared gain of the A.R. response function. The models are designated as Q,R and S.

MODEL Q

$$(1-0.8L)\, y(t) = -0.3\, x(t) + (1+0.5L)\, e(t)$$

MODEL R

$$(1-1.031L + 0.630L^2 - 0.524L^3)\, y(t) = -0.375\, x(t) + (1+0.5L)\, e(t)$$

MODEL S

$$(1-1.295L + 0.428L^2)\, y(t) = 1.0\, x(t) + (1+0.7L)\, e(t)$$

Models R and S will be familiar from Chapter 3 and as Model Q corresponds to a Koyck lag all are fairly typical of econometric models. The exogenous variable was generated by Scheme A(i) for Models Q and R with variance ratios of 100/1 while Model S has Scheme B(i) with variance ratio 1/1.

The summary statistics will be presented in a similar fashion to the previous chapter except that there will now be two parts: Part A will present the means, biases and standard errors of bias and Part B will give the sample and asymptotic standard deviations and the root mean square error (RMSE). Tables 4.1, 4.2, and 4.3 contain these statistics for both estimators for the above three models.

Table 4.1

Summary Statistics for A.T.D. and PH

Model Q: All Sample Sizes

PART A.

N	P	A.T.D.			PH		
		M	B	B/SE	M	B	B/SE
40	0.8	0.718	-0.082	5.22	0.715	-0.085	5.12
	-0.3	-0.486	-0.186	4.29	-0.488	-0.188	4.23
	0.5	0.548	0.048	1.75	0.546	0.046	1.71
70	0.8	0.760	-0.040	3.96	0.757	-0.043	4.26
	-0.3	-0.385	-0.085	4.34	-0.392	-0.092	4.60
	0.5	0.534	0.034	1.82	0.537	0.037	2.14
100	0.8	0.767	-0.033	3.79	0.768	-0.032	3.68
	-0.3	-0.367	-0.067	4.01	-0.367	-0.067	3.94
	0.5	0.537	0.037	2.70	0.533	0.033	2.46

PART B.

N	P	SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
40	AR	0.110	1.15	0.137	0.116	1.21	0.144
	EV	0.304	1.83	0.356	0.311	1.88	0.363
	MA	0.192	1.28	0.198	0.188	1.26	0.194
70	AR	0.071	1.01	0.081	0.071	1.01	0.083
	EV	0.137	1.09	0.161	0.140	1.11	0.168
	MA	0.131	1.13	0.135	0.121	1.04	0.126
100	AR	0.061	1.03	0.069	0.061	1.03	0.069
	EV	0.117	1.11	0.135	0.119	1.13	0.137
	MA	0.096	1.03	0.103	0.094	1.01	0.099

Table 4.2

Summary Statistics for A.T.D. and PH

Model R: All Sample Sizes

PART A

N	P	A.T.D.			PH		
		M	B	B/SE	M	B	B/SE
40	1.031	1.011	-0.020	0.58	0.995	-0.036	1.00
	-0.630	-0.673	-0.043	1.00	-0.654	-0.024	0.65
	0.524	0.523	-0.001	0.05	0.512	-0.012	0.55
	-0.375	-0.602	-0.227	3.47	-0.642	-0.267	3.57
	0.5	0.480	-0.020	0.46	0.499	-0.001	0.03
70	1.031	1.001	-0.030	1.16	0.995	-0.036	1.34
	-0.630	-0.611	0.019	0.58	-0.610	0.020	0.65
	0.524	0.500	-0.024	1.48	0.504	-0.020	1.29
	-0.375	-0.513	-0.138	3.49	-0.520	-0.145	3.61
	0.5	0.523	0.023	0.75	0.527	0.027	0.93
100	1.031	1.005	-0.026	1.72	1.008	-0.023	1.59
	-0.630	-0.594	0.036	1.64	-0.597	0.033	1.58
	0.524	0.505	-0.019	1.52	0.510	-0.014	1.14
	-0.375	-0.475	-0.100	3.19	-0.480	-0.105	3.34
	0.5	0.548	0.048	2.42	0.538	0.038	2.02

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
40	AR	0.2424	1.32	0.2432	0.2213	1.20	0.2242
	AR	0.3005	1.14	0.3036	0.2592	0.99	0.2603
	AR	0.1521	1.00	0.1521	0.1535	1.01	0.1540
	EV	0.4576	1.36	0.5108	0.5228	1.55	0.5870
	MA	0.3041	1.53	0.3048	0.2404	1.21	0.2404
70	AR	0.1803	1.31	0.1828	0.1875	1.36	0.1909
	AR	0.2288	1.17	0.2296	0.2163	1.10	0.2172
	AR	0.1137	1.01	0.1162	0.1082	0.96	0.1100
	EV	0.2762	1.14	0.3088	0.2817	1.16	0.3168
	MA	0.2153	1.45	0.2165	0.2022	1.36	0.2040
100	AR	0.1056	0.92	0.1088	0.1015	0.88	0.1041
	AR	0.1535	0.94	0.1577	0.1462	0.90	0.1499
	AR	0.0877	0.93	0.0897	0.0859	0.91	0.0870
	EV	0.2192	1.11	0.2409	0.2201	1.12	0.2439
	MA	0.1388	1.12	0.1469	0.1313	1.06	0.1367

Table 4.3

Summary Statistics for A.T.D. and PH

Model S: All Sample Sizes

PART A

N	P	A.T.D.			PH		
		M	B	B/SE	M	B	B/SE
40	1.295	1.281	-0.014	1.06	1.267	-0.028	1.67
	-0.428	-0.424	0.004	0.29	-0.414	0.014	0.79
	1.0	0.950	-0.050	2.09	0.952	-0.048	1.99
	0.7	0.753	0.053	2.15	0.773	0.073	2.77
70	1.295	1.295	0.000	0.00	1.275	-0.020	1.48
	-0.428	-0.416	0.012	0.86	-0.414	0.014	1.00
	1.0	0.971	-0.029	1.92	0.970	-0.030	2.03
	0.7	0.767	0.067	3.76	0.761	0.061	3.41
100	1.295	1.289	-0.006	0.51	1.289	-0.006	0.53
	-0.428	-0.425	0.003	0.25	-0.426	0.002	0.17
	1.0	0.984	-0.016	1.55	0.987	-0.013	1.24
	0.7	0.737	0.037	2.70	0.733	0.033	2.52

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
40	AR	0.0934	0.87	0.0944	0.1189	1.11	0.1222
	AR	0.0974	0.91	0.0975	0.1252	1.16	0.1260
	EV	0.1692	1.16	0.1764	0.1704	1.16	0.1770
	MA	0.1748	1.41	0.1827	0.1865	1.51	0.2003
70	AR	0.0963	1.20	0.0963	0.0953	1.19	0.0974
	AR	0.0973	1.21	0.0980	0.0987	1.23	0.0997
	EV	0.1054	0.96	0.1093	0.1047	0.96	0.1089
	MA	0.1246	1.35	0.1415	0.1266	1.37	0.1405
100	AR	0.0816	1.22	0.0818	0.0809	1.21	0.0811
	AR	0.0844	1.26	0.0844	0.0834	1.25	0.0834
	EV	0.0720	0.79	0.0738	0.0740	0.81	0.0751
	MA	0.0959	1.25	0.1028	0.0925	1.20	0.0982

Tables 4.1, 4.2 and 4.3 reveal a close agreement between the two estimators in all samples indicating that it matters little whether e* is treated as a constant or a stochastic variable. For samples of size 40 there are some differences and a formal test for equality of means and variances yielded the following t and F values:-

(a) The highest t value was 0.65 for β_1 of Model S.

(b) The highest F values were 1.59 (α_2 of Model R), 1.61 and 1.66 (β_1 and β_2 of Model S).

Only the F values are significant, and then only marginally, so that there is some support for the supposition that A.T.D. may be a slightly more efficient estimator in small samples.

Apart from the relative performance of the estimators it is interesting to observe the behaviour of each of the estimators individually. In particular the large bias on the exogenous variable noted in connection with the Nerlove filter in Chapter 3 is carried through here to A.T.D., thereby establishing that it is not peculiar to PH. Throughout the thesis there was an accumulation of evidence that the nature of the exogenous regressors was a determinant of the small sample performance of ARMAX estimators. However it is very difficult to design a controlled experiment to measure this influence as changing filter level and shape both affect the "signal/noise" ratio (R^2). Nevertheless it is interesting to note the behaviour of PH in the estimation of Models Q and F when the variance ratios are not 100/1 and 500/1 as in Tables 4.1 and 3.7 (p. 126) and accordingly reference is made to Table 4.4.

Table 4.4

Summary Statistics for Models Q and F

Sample Size N = 100 : Various Variance Ratios

Model	Variance Ratio	P	M	B	SE	B/SE
Q	1/5	0.8	0.8003	0.0003	0.0008	0.38
		-0.3	-0.3001	-0.0001	0.0012	0.08
		0.5	0.5119	0.0119	0.0131	0.91
F	5/1	1.031	1.0821	0.0511	0.0239	2.14
		-0.630	-0.6888	-0.0588	0.0257	2.29
		0.524	0.5350	0.0110	0.0104	1.06
		-0.375	-0.3657	0.0093	0.0068	1.37
		0.8	0.7607	-0.0393	0.0369	1.07
		0.3	0.2782	-0.0218	0.0342	0.64

With smaller variance ratios the bias on the exogenous variable has all but disappeared, although there appears to have been an associated rise in the biases of the A.R. parameters for Model F. However the degree of emphasis to be given to the results of Table 4.4 is uncertain as can be seen from viewing the O.L.S. estimates of 0.8001, -0.2990 i.e. the signal is so powerful as to swamp any noise and the computed R^2 is boosted to around 0.999. Therefore the relative roles of filter shape and level are not distinct and a proper evaluation would require a great many more experiments with differing variance ratios. In performing such experiments it would be necessary to decide if R^2 should be kept in a "realistic" range or allowed to become very high - as in the experiments of Table 4.4 - bearing in mind that any decision to limit it would restrict the permissible variance ratios quite severely. No attempt was made to persue this line of enquiry owing to the demands of computer time but it is apparent that the nature of exogenous regressors is a

factor that should be analysed in much greater detail.[1]

As the sample size changes the M.A. parameters vary in an odd way. In every instance the bias is significant in sample size 100 but not for the two smaller samples. To explain such behaviour for Models R and S we note the heavy enforcement of the root restrictions in the smaller samples : operating on the M.A. for Model S and the A.R. for Model R. No such explanation is available for Model Q, but as it possesses identical A.R. and M.A. transfer functions to Model A, and it will be recalled that PH performed poorly on the latter model (see p. 119), the sampling results of Table 4.3 are scarcely surprising.

All sampling variances are not significantly different from their asymptotic values in samples of 70, so that as the sampling performance of A.T.D. and PH was close for all models the choice of estimator must be a function of other factors, the principal one being computational cost. Table 4.5, in presenting the C.P.U. time taken for 50 replications of Model R, provides some information on this.

Table 4.5

Computation Time for A.T.D. and PH for Model R

All Sample Sizes

N	Estimator	C.P.U. time (mins.)
40	A.T.D.	7.14
	PH	6.82
70	A.T.D.	7.77
	PH	8.92
100	A.T.D.	10.05
	PH	9.98

1 Some support for the level hypothesis might be gleaned from the H-T study in the form of the significant coefficient attached to the "signal/noise" ratio in the bias regressions. Unfortunately, as argued in Appendix 7, any interpretation of the parameters of these regressions is not unique.

Computationally, also, there appears to be little to choose between the estimators. Although this conclusion is true for Models Q,R and S (Table 4.5 is a fair representation of the rankings for the others) it was found to be misleading as the order of the M.A. increased - a result of the differing methods of inversion possible with each estimator. This is best illustrated by inspecting the form of the first derivatives of the sum of squares function S with respect to the j'th parameter in the vector of A.R. and E.V. parameters (see equations (2.14) and (2.17)):

$$\frac{\partial S}{\partial \delta_j} = -2Z_j^T (M^T M)^{-1} e \qquad \text{(PH)}$$

$$\frac{\partial S}{\partial \delta_j} = -2Z_j^T \Omega^{-1} u \qquad \text{(A.T.D.).}$$

As M is a triangular matrix it is possible to form $\frac{\partial S}{\partial \delta_j}$ for PH purely by one application of recursion relations - a very fast procedure for a computer - but such is not the case for A.T.D. as Ω must be inverted and it is NxN. Although there is some computational gain from the band structure of Ω, for an ARMA (p,q) model the number of multiplications performed in the inversion is of the order of q for PH and q^2 for A.T.D. Clearly for the high order M.A. processes found in some monthly models, unless some efficient method of inverting band matrices can be found, A.T.D. would not be feasible.

4.3 A Comparison of PH/0 and PH

Section 2.5.1 has demonstrated that asymptotically PH/0 and PH should be equivalent but there is a possibility that in small samples the saving of q degrees of freedom will be sufficient to give a slight advantage to PH/0. For high order models e.g. $(1 + \alpha_{12}L^{12})^2$ it seems

doubtful that the reduction in the sum of squares from fitting all e* would ever justify such a step, but for the case of most interest to this thesis i.e. quarterly data there are unlikely to be more than fifth order M.A. terms.

Three models are used for the comparison - Models E and F of Chapter 3 and Model T below.

MODEL T

$y(t) = (1+0.2L)\ (1+0.4L^4)\ e(t)$.

Appendix 8 contains the spectral shape of Model T and this is reminiscent of a quarterly time series (with moderate trend) or the error structure arising from a quarterly series that has undergone differencing transformations. Tables 4.6, 4.7 and 4.8 present the summary statistics for the two estimators for all three sample sizes.

Table 4.6

Summary Statistics for PH/O and PH

Model E,F and T : Sample Size N = 40

PART A

Model	P	PH/O			PH		
		M	B	B/SE	M	B	B/SE
T	0.2	0.2001	0.0001	0.01	0.1869	-0.0131	0.53
	0.4	0.4058	0.0058	0.25	0.5261	0.1261	3.45
E	1.55	1.4947	-0.0553	2.04	1.4250	-0.1250	3.46
	-0.6	-0.5629	0.0371	1.39	-0.4976	0.1024	2.92
	0.8	0.8499	0.0499	1.42	0.9320	0.1320	2.82
	0.3	0.3463	0.0463	1.29	0.4514	0.1514	3.33
F	1.031	1.1351	0.1041	1.81	1.0082	-0.0228	0.46
	-0.630	-0.7785	-0.1485	2.90	-0.6969	-0.0669	1.29
	0.524	0.5097	-0.0143	0.56	0.5379	0.0139	0.62
	-0.375	-0.6218	-0.2468	2.44	-0.6405	-0.2655	2.40
	0.8	0.6298	-0.1702	2.20	0.7389	-0.0611	0.85
	0.3	0.1895	-0.1105	1.58	0.3548	0.0548	0.96

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
T	MA	0.1684	1.10	0.1684	0.1750	1.15	0.1755
	MA	0.1661	1.16	0.1662	0.2582	1.80	0.2873
E	AR	0.1956	1.10	0.2033	0.2553	1.44	0.2825
	AR	0.1888	1.08	0.1924	0.2485	1.42	0.2688
	MA	0.2488	1.21	0.2538	0.3310	1.61	0.3563
	MA	0.2539	1.30	0.2581	0.3218	1.65	0.3556
F	AR	0.4064	1.45	0.4195	0.3525	1.26	0.3532
	AR	0.3622	1.08	0.3915	0.3664	1.09	0.3725
	AR	0.1799	1.12	0.1805	0.1582	0.99	0.1588
	EV	0.7163	1.75	0.7576	0.7828	1.91	0.8266
	MA	0.5461	1.75	0.5720	0.5060	1.62	0.5097
	MA	0.4953	1.82	0.5075	0.4016	1.48	0.4053

Table 4.7

Summary Statistics for PH/O and PH

Models E,F and T : Sample Size N = 70

PART A

Model	P	PH/O			PH		
		M	B	B/SE	M	B	B/SE
T	0.2	0.2206	0.0206	1.23	0.2180	0.0180	1.06
	0.4	0.3810	-0.0190	1.28	0.4178	0.0178	0.90
E	1.55	1.5454	-0.0046	0.27	1.5255	-0.0245	1.23
	-0.6	-0.6101	-0.0101	0.59	-0.5900	0.0100	0.50
	0.8	0.8010	0.0010	0.04	0.8380	0.0380	1.48
	0.3	0.2904	-0.0096	0.42	0.3211	0.0211	0.84
F	1.031	1.0770	0.0460	1.13	1.0056	-0.0254	0.58
	-0.630	-0.6902	-0.0602	1.56	-0.6225	0.0075	0.17
	0.524	0.5038	-0.0202	1.21	0.4974	-0.0266	1.45
	-0.375	-0.5082	-0.1332	2.21	-0.5691	-0.1941	2.58
	0.8	0.7305	-0.0695	1.38	0.8253	0.0253	0.47
	0.3	0.2402	-0.0598	1.37	0.3256	0.0256	0.56

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
T	MA	0.1184	1.03	0.1202	0.1204	1.04	0.1217
	MA	0.1057	0.98	0.1074	0.1399	1.29	0.1410
E	AR	0.1224	0.91	0.1225	0.1410	1.05	0.1431
	AR	0.1214	0.92	0.1218	0.1428	1.08	0.1431
	MA	0.1704	1.10	0.1704	0.1816	1.17	0.1855
	MA	0.1615	1.09	0.1618	0.1765	1.20	0.1778
F	AR	0.2871	1.36	0.2908	0.3105	1.47	0.3115
	AR	0.2721	1.07	0.2787	0.3173	1.25	0.3174
	AR	0.1179	0.98	0.1196	0.1294	1.07	0.1321
	EV	0.4257	1.36	0.4461	0.5322	1.71	0.5665
	MA	0.3550	1.51	0.3617	0.3803	1.62	0.3811
	MA	0.3090	1.51	0.3147	0.3250	1.59	0.3260

Table 4.8

Summary Statistics for PH/O and PH

Models E,F and T: Sample Size N = 100

PART A

Model	P	PH/O			PH		
		M	B	B/SE	M	B	B/SE
T	0.2	0.2089	0.0089	0.65	0.2106	0.0106	0.75
	0.4	0.3667	-0.0333	2.71	0.3919	-0.0081	0.60
E	1.55	1.5541	0.0041	0.27	1.5369	-0.0131	0.81
	-0.6	-0.6172	-0.0172	1.21	-0.6007	-0.0007	0.05
	0.8	0.7877	-0.0123	0.67	0.8178	0.0178	0.89
	0.3	0.2703	-0.0297	1.59	0.2914	-0.0086	0.43
F	1.031	1.0514	0.0204	0.85	0.9845	-0.0465	1.76
	-0.630	-0.6635	-0.0335	1.31	-0.5846	0.0454	1.48
	0.524	0.5155	-0.0085	0.64	0.5061	-0.0179	1.28
	-0.375	-0.4855	-0.1105	2.55	-0.5157	-0.1407	2.90
	0.8	0.7642	-0.0358	1.32	0.8483	0.0483	1.44
	0.3	0.2541	-0.0459	1.63	0.3185	0.0185	0.57

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
T	MA	0.0961	1.00	0.0965	0.1002	1.04	0.1007
	MA	0.0872	0.96	0.0933	0.0963	1.06	0.0966
E	AR	0.1067	0.95	0.1068	0.1144	1.02	0.1151
	AR	0.1001	0.90	0.1016	0.1095	0.99	0.1095
	MA	0.1304	1.00	0.1310	0.1406	1.08	0.1417
	MA	0.1325	1.07	0.1358	0.1428	1.16	0.1431
F	AR	0.1702	0.96	0.1714	0.1869	1.06	0.1926
	AR	0.1813	0.86	0.1844	0.2173	1.03	0.2220
	AR	0.0932	0.92	0.0936	0.0991	0.98	0.1007
	EV	0.3069	1.19	0.3262	0.3433	1.33	0.3710
	MA	0.1923	0.98	0.1956	0.2367	1.20	0.2416
	MA	0.1989	1.16	0.2041	0.2312	1.35	0.2319

Tables 4.6, 4.7 and 4.8 illustrate the following features of the estimators:

(i) PH/0 is a more efficient estimator than PH as a consequence of the restriction e* = 0 imposed by the former. There is an exception to the conclusion in Model F when N = 40, probably because the root restrictions were enforced at 16 replications for PH and at only 4 for PH/0. That enforcement is a necessary, but certainly not a sufficient condition for a lower variance, may be seen in the example of Model T (Table 4.9 p.175).

(ii) Generally biases are similar for both estimators with the only exceptions being Model T (N = 100) and all models when N = 40. The large bias associated with α_4 in Model T for PH/0 when N = 100 is peculiar and it was not possible to advance any explanation of this. In contrast there is the large bias for the PH estimator of this parameter in the smallest sample and a lack of it for PH/0. Just over one half of the discrepancy is due to eight replications which yielded coefficient estimates on the unit circle for PH and these are presented in Table 4.9.

Why should PH give unit roots while PH/0 does not? A number of answers were canvassed. The first was that the behaviour is unique to PH, but this is easily refuted as both B-J and A.T.D. gave the same results (for B-J see Table 4.16).[2] A second explanation recognizes that although the distribution of the estimator must lie between -1 and 1 the assumption of normality gives a range of $(-\infty,\infty)$, so there is a finite

---

2 Strangely A.F.D. does not give unit roots and corresponds more closely to PH/0 than the other estimators. This will be mentioned later.

Table 4.9

Replications for which the Root Restrictions are Enforced

Model T : N = 40

Replication No.	PH/0	PH
3	0.77	1.00
4	0.90	1.00
18	0.44	1.00
27	0.56	1.00
32	0.64	1.00
43	0.51	1.00
45	0.51	1.00
47	0.49	1.00

probability of any estimate being outside the unit circle at a given replication. A measure of the magnitude of this probability comes from the (approximate) asymptotic standard deviation $\sigma_{\hat{\alpha}_4} = N^{-1}(1-\alpha_4^2)^{\frac{1}{2}}$, where the true value of α_4 is inserted. Then (asymptotically) there is a 95% chance that the estimate will lie in the range $\alpha_4 \pm 1.96\ \sigma_{\hat{\alpha}_4}$.

Evidence to support the hypothesis that the unit roots come from the tail of a continuous distribution may be deduced by observing that because $\sigma_{\hat{\alpha}_4}$ is a function of N, as $N \to \infty$ the probability of $\hat{\alpha}_4$ coinciding with the boundary becomes smaller, and when N equalled 70 only replication 4 displayed any perverse behaviour. For N = 100 no instances were recorded. Contrary to this is the fact that, even if the extreme viewpoint of the asymptotic variances understating the small sample variances by a factor of two (not impossible) is adopted, it would not be sufficient to explain 8 out 50 estimates being greater than unity. At best it might explain one.

Therefore there doesn't appear to be anything in the nature of the estimators that would provide a simple explanation of the phenomenon. As a next step it is natural to scrutinize the data for any clues and this is done most conveniently by inspecting the correlogram of y(t) for the offending replications. As a help in analysing the correlogram we note two characteristics that are potentially associated with unit roots in the M.A.

(i) As with the variance of the estimator it is possible that the simulated data has thrown up empirical covariances consistent with unit roots. An example of this occurs in Wold's analysis of Beveridge's wheat yield data [119 p.154] in which the first serial correlation coefficient ($\hat{\rho}_1$) was greater than 0.5, so that fitting a first order M.A. by solving the covariance equations leads to no $\hat{\alpha}_1 < 1.0$. In the same way there is a finite probability that $\hat{\rho}_4$ might be greater than 0.5 (its theoretical value when $\hat{\alpha}_1 = 0.2$, $\hat{\alpha}_4 = 1.0$). Asymptotically the mean of the distribution of $\hat{\rho}_4$ is 0.3448 and its variance is approximately $\frac{1}{N}$ so that there is a 35% chance at any replication that $\hat{\rho}_4$ will be greater than 0.5. In fact only three of the replications had $\hat{\rho}_4$ greater than 0.5 and the mean and standard deviation over fifty replications were 0.2864 and 0.1268, which differ substantially from the asymptotic values.[3]

(ii) As $\hat{\alpha}_4 \to 1$ the series y(t) should possess a correlogram reminiscent

---

[3] In part the discrepency between 35% (expected) and 16% (observed) is due to the critical size selected for $\hat{\rho}_4$. This value is conditional on $\hat{\alpha}_1 = 0.2$ so that, if at any replication $\hat{\alpha}_1 \neq 0.2$, it is possible for the critical value of $\hat{\rho}_4$ to exceed 0.5. No conclusions in the text were altered by adjusting for this.

of a non-stationary series i.e. slowly declining. Table 4.11 presents the correlogram of the first six lags for the replications of interest.

Table 4.11

Correlogram of y(t) for Replications 3,4,18,27,32,43,45 and 47

Model T: N = 40

Replication No.	Lag					
	1	2	3	4	5	6
3	0.018	-0.319	-0.140	0.402	0.033	-0,151
4	0.212	-0.105	0.111	0.416	-0.086	0.060
18	0.126	-0.395	-0.051	0.215	-0.121	-0.219
27	-0.043	0.053	0.112	0.415	0.083	0.044
32	-0.125	0.047	-0.110	0.252	-0.160	-0.009
43	0.056	-0.074	-0.001	0.402	0.116	-0.197
45	0.080	-0.015	0.068	0.351	-0.181	-0.133
47	0.056	-0.412	-0.526	0.560	0.066	-0.228

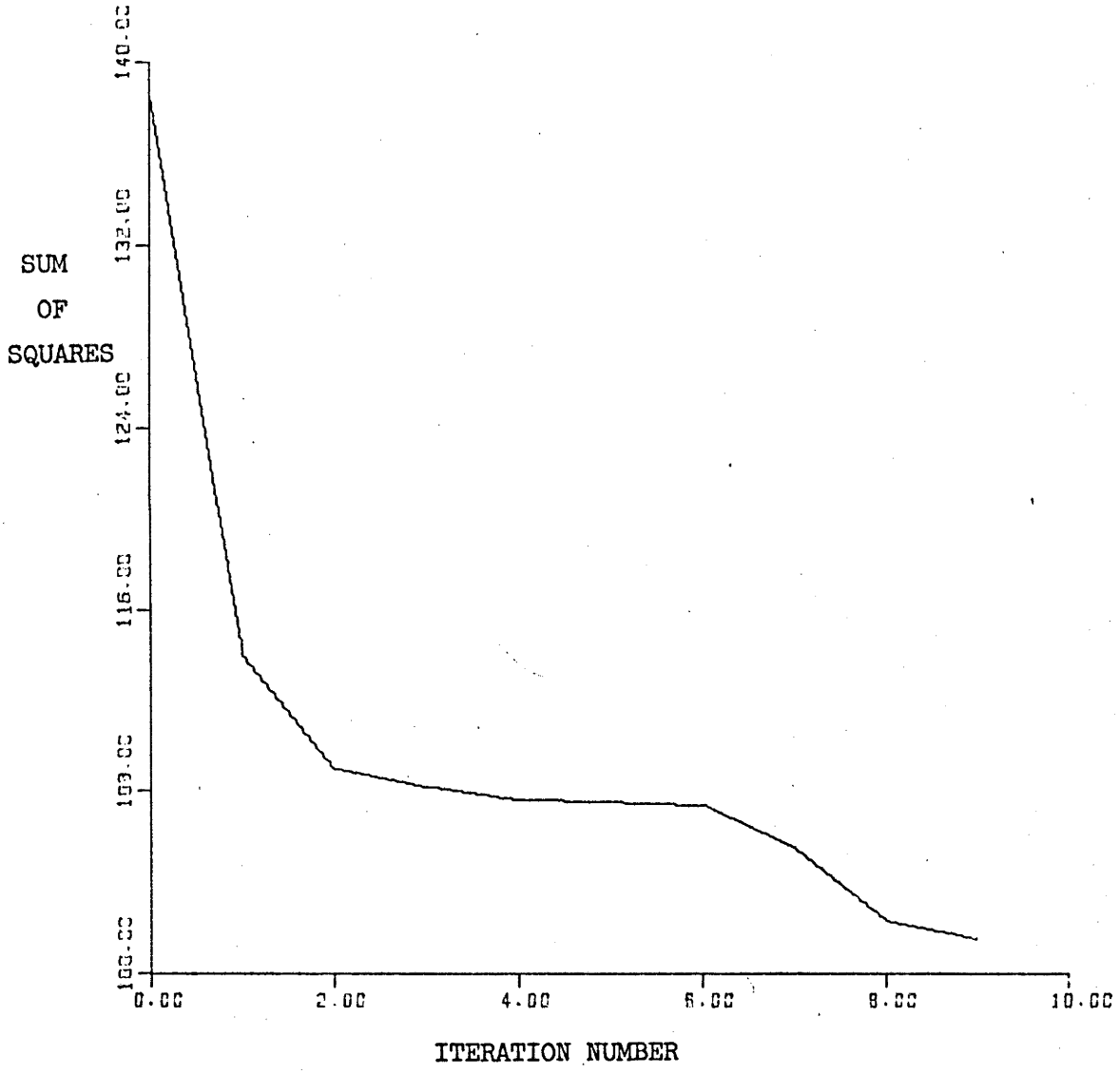
There is no evidence in any of the correlograms to support the hypothesis of a non-stationary series. In addition the estimated residual variance of the errors was always close to 3.0 (the population value) for these replications leading to the conclusion that the estimates of e* are dominating the series i.e. the transient introduced by their inclusion is more important than the other contributions. This concurs with the improved behaviour as the sample grows in that the transient will have a progressively smaller impact upon the solution to the difference equation.

A final view of the problem is contained in the sum of squares function for replication 26. In Fig. 4.1 the sum of squares function is plotted against $\hat{\alpha}_4$ for the 10 replications taken for convergence from the starting value of 0.4 to the boundary of unity. The pattern that emerges is typical of those replications in which the root restrictions were enforced, in that a plateau seems to be defined around 0.75 and

FIGURE 4.1

SUM OF SQUARES FUNCTION FOR MODEL T

AS ITERATION PROGRESSES



it appears that this will be the final value, but then there is a sharp change in the parameters and (toward the end) the sum of squares.[4] Again this supports the contention that the estimated e* are unduly influencing the sum of squares but it should be remembered that A.T.D. followed the same pattern.

Although a recitation of the features of the data from each of the replications has been given no solutions have emerged, making it necessary to consider whether the root restrictions should be incorporated into the estimation procedure in a more rigorous manner. Corresponding to the dichotomy in the theory of statistical inference there could be two approaches to this.

(a) In a Bayesian framework the restrictions would be imposed through the prior distribution of the M.A. parameters. This seems to have been done successfully by Zellner and Geisel [125] in the context of a first order A.R. in the disturbances but obviously the numerical integration required will become exceedingly complex as the number of parameters grow, and it is dubious whether the technique can become of practical use in the face of the large number of parameters found in many econometric models. Additionally there is the complication that the

---

[4] Struik has found this for the Box and Jenkins' airline data as well (when only 9 years of data were used) and he comments that the ".... minimum of the sum of squares occurred in the vicinity of the boundary....but that there was some suggestion of an inflection of the surface in the neighbourhood of $\alpha_1 = 0.4$, $\alpha_4 = 0.6$." [102 p.14].

analytical relationship between root and parameter restrictions is not easy to determine once the order of polynomials becomes greater than two.

(b) The classical approach would center around Barnard's mean likelihood estimator [6]. Jenkins and Watts [55 p.194-195] apply this to first and second order A.R.'s, and for the first order case give a numerical example in which the M.L. estimate of 0.9 is changed to 0.86. Unfortunately, when there are n parameters, two n-fold integrations must be performed with the limits of integration being determined by the boundary condition so that this solution has the same disadvantages as (a) above.

Our conclusion must be pessimistic. There seems no good reason for the odd behaviour of PH, A.T.D. and B-J and there are no simple modifications to guard against it. The strange behaviour of the estimators must make them suspect until future research can provide some rationale for the observed patterns.

Having strayed from the PH/O, PH comparison it is now time to return in order to discuss the computational advantages of the former estimator. Table 4.12 contains the C.P.U. time for 50 replications of Models E and T.

Table 4.12

Computation Time for PH/0 and PH

Models E and T : All Sample Sizes

Model	Sample	C.P.U. Time (Mins.)	
		PH/0	PH
E	40	2.72	5.03
	70	2.85	5.01
	100	3.47	6.00
T	40	1.06	5.94
	70	1.69	6.17
	100	1.86	8.05

From Table 4.12 is is apparent that PH/0 is a faster algorithm than PH. Most of the speed is gained from the smaller number of iterations necessary to reach the minimum, and as a general principle the number of iterations was related to the number of parameters to be estimated. Allied with Tables 4.6 - 4.8, Table 4.12 provides a powerful incentive for the use of PH/0 in almost all sample sizes that are found with quarterly time series and certainly justifies its adoption for model specification. Although the roots in the M.A. of the models were not high - and therefore PH/0 should perform well - the strange results obtained with PH when the roots are large,leaves it open to doubt as well.

4.4 A Comparison of O.L.S. and PH

As shown in Chapter 1 O.L.S. has a number of non-optimal properties in large samples when a M.A. is present in the disturbance process viz. loss of efficiency when the equation was ARMAX (o,r,q) and inconsistency when ARMAX (p,r,q), and an evaluation of the size of each indicated that these could be quite substantial. For small samples the

position is not so clear, in that the combination of a considerable gain in degrees of freedom and the powerful minimum variance tendency of O.L.S. found by other researchers, may react to produce a smaller RMSE than the M.L. estimator.

To assess the small sample effects three models were chosen - Models A, D and G. The basis for selection of the first two was the relatively poor performance of PH in their estimation so that the comparison will be favourable to O.L.S., while the last model represents a more realistic case. Tables 4.13, 4.14 and 4.15 contain the summary statistics.

Table 4.13

Summary Statistics for O.L.S. and PH

Models A, D and G : Sample Size N = 40

PART A

Model	P	O.L.S.			PH		
		M	B	B/SE	M	B	B/SE
A	0.8	0.8550	0.0550	5.98	0.741	-0.059	3.28
D	1.031	1.4941	0.4631	25.87	0.9388	-0.0922	1.97
	-0.630	-1.1352	-0.5052	30.07	-0.6037	0.0263	0.50
	0.524	0.5882	0.0642	4.69	0.5217	-0.0023	0.09
G	1.295	1.4522	0.1572	11.39	1.2377	-0.0573	2.40
	-0.428	-0.5825	-0.1545	12.98	-0.4012	0.0268	1.28
	1.0	0.9077	-0.0923	2.82	1.1690	0.1690	3.61

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
A	AR	0.0648	0.60	0.0850	0.1264	1.21	0.1395
D	AR	0.1262	0.45	0.4800	0.3312	1.18	0.3438
	AR	0.1188	0.35	0.5190	0.3722	1.10	0.3731
	AR	0.0970	0.60	0.1163	0.1763	1.09	0.1763
G	AR	0.0977	0.57	0.1850	0.1691	0.98	0.1785
	AR	0.0841	0.56	0.1759	0.1475	0.97	0.1499
	EV	0.2315	0.76	0.2492	0.3310	1.08	0.3716

Table 4.14

Summary Statistics for O.L.S. and PH

Models A, D and G : Sample Size N = 70

PART A

Model	P	O.L.S.			PH		
		M	B	B/SE	M	B	B/SE
A	0.8	0.8662	0.0662	10.18	0.759	-0.041	3.18
D	1.031	1.5263	0.4953	33.24	0.9590	-0.0720	2.15
	-0.630	-1.1510	-0.5210	36.43	-0.5610	0.0690	1.76
	0.524	0.5748	0.0508	4.66	0.4737	-0.0503	2.09
G	1.295	1.4849	0.1899	17.26	1.2674	-0.0276	1.39
	-0.428	-0.6054	-0.1774	18.87	-0.4213	0.0067	0.40
	1.0	0.8548	-0.1452	5.69	1.1183	0.1183	3.23

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
A	AR	0.0461	0.59	0.0807	0.0905	1.15	0.0993
D	AR	0.1051	0.49	0.5063	0.2371	1.11	0.2478
	AR	0.1011	0.40	0.5307	0.2766	1.09	0.2851
	AR	0.0773	0.63	0.0925	0.1512	1.23	0.1593
G	AR	0.0775	0.60	0.2051	0.1406	1.08	0.1433
	AR	0.0666	0.58	0.1895	0.1199	1.05	0.1200
	EV	0.1806	0.79	0.2317	0.2585	1.13	0.2843

Table 4.15

Summary Statistics for O.L.S. and PH

Models A, D and G : Sample Size N = 100

PART A

Model	P	O.L.S.			PH		
		M	B	B/SE	M	B	B/SE
A	0.8	0.8680	0.0680	11.53	0.762	-0.038	3.49
D	1.031	1.5253	0.4943	42.25	0.9563	-0.0747	2.81
	-0.630	-1.1472	-0.5172	46.18	-0.5541	0.0759	2.40
	0.524	0.5762	0.0522	6.21	0.4799	-0.0441	2.96
G	1.295	1.4900	0.1950	20.97	1.2697	-0.0253	1.55
	-0.428	-0.6088	-0.1808	22.32	-0.4219	0.0061	0.44
	1.0	0.8431	-0.1569	7.30	1.1090	0.1090	3.59

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
A	AR	0.0414	0.63	0.0796	0.0763	1.17	0.0852
D	AR	0.0828	0.46	0.5012	0.1880	1.05	0.2023
	AR	0.0792	0.37	0.5232	0.2232	1.05	0.2357
	AR	0.0592	0.57	0.0789	0.1053	1.02	0.1142
G	AR	0.0660	0.61	0.2059	0.1154	1.06	0.1181
	AR	0.0575	0.60	0.1897	0.0993	1.04	0.0995
	EV	0.1521	0.79	0.2185	0.2149	1.12	0.2410

It is a relatively simple task to draw conclusions from Tables 4.13-4.15 as the summary statistics retain the same rankings for all sample sizes. The principal items of interest are:

(a) The powerful minimum variance property of O.L.S. is in evidence for all samples and models and understates the variance of the M.L. estimator by a factor in the vicinity of two. This leads to a smaller RMSE for O.L.S. than PH for a number of parameters - β_1 of Model A, β_3 of Model D, and γ_0 of Model G but in the light of the fact that the

first two models (and γ_0 of Model G) were not estimated very accurately by PH this is hardly surprising. In Section 1.5 Model A was featured in deriving the asymptotic inconsistency and formula variance of O.L.S. When N = 100 the theoretical inconsistency is 0.085 and the efficacy ratio is 2.15. Both of these values are broadly in agreement with Table 4.15 implying that the formula and actual variance of O.L.S. may be much closer in ARMA models than was the case when exogenous variables were present.

(b) The biases in O.L.S. are very significant and show little variation with sample size. Moreover the biases tend to cancel, thereby having the effect that although the individual parameters are poorly estimated their sum is not, and this may mean that statistics such as the average lag are robust against autocorrelation.[5]

(c) To some extent the models selected are favourable to O.L.S. This can be explained by observing that the M.L. estimator minimizes

$$S = \sum_j \frac{I_{yy}(\lambda_j)|B(\lambda_j)|^2}{|A(\lambda_j)|^2} ,$$

while O.L.S. minimizes

$$S = \sum_j I_{yy}(\lambda_j)|B(\lambda_j)|^2.$$

The relationship of the O.L.S. estimate to the M.L. estimate will then be a function of

(i) The flatness of the disturbance spectrum i.e. $|A(\lambda)|^2$. If the disturbance is white noise the spectrum is flat and O.L.S. is fully

---

[5] The conclusion concerning the sum held for all models fitted by O.L.S.

efficient as equal weighting is then given to each of the disturbance periodogram ordinates. If the spectrum is not flat an equal weighting will be incorrect. For the simulated models of this section the ratio of $|A(\lambda)|^2$ at frequencies 0 and π are 9/1 (Models A and G) and 17/1 (Model D), which is not a great departure from a uniform spectrum.

(ii) The shape of $I_{yy}(\lambda)|B(\lambda)|^2$ e.g. for Model G this expression has most of its power concentrated near the origin so that it is only necessary to discover whether $|A(\lambda)|^2$ is flat over a relatively small band near the origin and, this being the case, O.L.S. will be close to the M.L. estimate. When there are peaks in the disturbance spectrum (say at $\pi/2$) O.L.S. will be considerably worse than the M.L. estimator (see Section 1.5), prompting the conclusion that the smooth monotonically declining disturbance spectra used in the experiments above will result in a bias in favour of O.L.S.

4.5 A Comparison of B-J and PH

These estimators were discussed in Section 2.5.4 and an evaluation of each in the context of a first order M.A. was given. In this section a number of models of greater complexity will be simulated viz. Models E,F and T. The choice was conditioned by the familiar spectral properties of E and F and the need to study the perverse behaviour of B-J in estimating Model T for N = 40. Initially one supposition was that if it was the estimation of e* that was accountable for the trait it might be avoided by an inclusion of e* in the objective function minimized i.e. it was postulated that because (for PH) S was only an indirect function of e* these estimates might be the cause of the sum of squares function having a minimum at the boundary. Tables 4.16, 4.17 and 4.18 contain the results for all three models.

Table 4.16

Summary Statistics for B-J and PH

Models E, F and T : Sample Size N = 40

PART A

Model	P	B-J			PH		
		M	B	B/SE	M	B	B/SE
E	1.55	1.4346	-0.1154	3.82	1.4250	-0.1250	3.46
	-0.6	-0.5065	0.0935	3.20	-0.4976	0.1024	2.92
	0.8	0.9339	0.1339	3.38	0.9320	0.1320	2.82
	0.3	0.4549	0.1549	3.74	0.4514	0.1514	3.33
F	1.031	0.9800	-0.0510	0.94	1.0082	-0.0228	0.46
	-0.630	-0.6507	-0.0207	0.38	-0.6969	-0.0669	1.29
	0.524	0.5119	-0.0121	0.46	0.5379	0.0139	0.62
	-0.375	-0.6962	-0.3212	2.85	-0.6405	-0.2655	2.40
	0.8	0.8160	0.0160	0.22	0.7389	-0.0611	0.85
	0.3	0.3786	0.0786	1.19	0.3548	0.0548	0.96
T	0.2	0.2103	0.0103	0.42	0.1869	-0.0131	0.53
	0.4	0.5013	0.1013	2.98	0.5261	0.1261	3.45

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
E	AR	0.2136	1.21	0.2428	0.2533	1.44	0.2825
	AR	0.2063	1.18	0.2265	0.2485	1.42	0.2688
	MA	0.2801	1.36	0.3104	0.3310	1.61	0.3563
	MA	0.2929	1.50	0.3313	0.3218	1.65	0.3556
F	AR	0.3832	1.37	0.3866	0.3525	1.26	0.3532
	AR	0.3905	1.16	0.3910	0.3664	1.09	0.3725
	AR	0.1862	1.16	0.1866	0.1582	0.99	0.1588
	EV	0.7980	1.95	0.8602	0.7828	1.91	0.8266
	MA	0.5174	1.66	0.5176	0.5060	1.62	0.5097
	MA	0.4662	1.72	0.4728	0.4016	1.48	0.4053
T	MA	0.1737	1.14	0.1740	0.1750	1.15	0.1755
	MA	0.2403	1.68	0.2608	0.2582	1.80	0.2873

Table 4.17

Summary Statistics for B-J and PH

Models E, F and T : Sample Size N = 70

PART A

Model	P	B-J			PH		
		M	B	B/SE	M	B	B/SE
E	1.55	1.5252	-0.0248	1.32	1.5255	-0.0245	1.23
	-0.6	-0.5897	0.0103	0.54	-0.5900	0.0100	0.50
	0.8	0.8355	0.0355	1.39	0.8380	0.0380	1.48
	0.3	0.3185	0.0185	0.75	0.3211	0.0211	0.84
F	1.031	1.0174	-0.0136	0.33	1.0056	-0.0254	0.58
	-0.630	-0.6325	-0.0025	0.06	-0.6225	0.0075	0.17
	0.524	0.4984	-0.0256	1.50	0.4974	-0.0266	1.45
	-0.375	-0.5433	-0.1683	2.56	-0.5691	-0.1941	2.58
	0.8	0.8084	0.0084	0.17	0.8253	0.0253	0.47
	0.3	0.3060	0.0060	0.14	0.3256	0.0256	0.56
T	0.2	0.2202	0.0202	1.19	0.2180	0.0180	1.06
	0.4	0.4119	0.0119	0.70	0.4178	0.0178	0.90

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
E	AR	0.1333	1.00	0.1355	0.1410	1.05	0.1431
	AR	0.1352	1.02	0.1355	0.1428	1.08	0.1431
	MA	0.1801	1.16	0.1835	0.1816	1.17	0.1855
	MA	0.1742	1.18	0.1752	0.1765	1.20	0.1778
F	AR	0.2888	1.37	0.2891	0.3105	1.47	0.3115
	AR	0.3057	1.21	0.3057	0.3173	1.25	0.3174
	AR	0.1207	1.00	0.1233	0.1294	1.07	0.1321
	EV	0.4652	1.49	0.4947	0.5322	1.71	0.5665
	MA	0.3566	1.52	0.3567	0.3803	1.62	0.3811
	MA	0.3074	1.50	0.3074	0.3250	1.59	0.3260
T	MA	0.1202	1.04	0.1218	0.1204	1.04	0.1217
	MA	0.1194	1.10	0.1200	0.1399	1.29	0.1410

Table 4.18

Summary Statistics for B-J and PH

Models E, F and T : Sample Size N = 100

PART A

Model	P	B-J			PH		
		M	B	B/SE	M	B	B/SE
E	1.55	1.5408	-0.0092	0.58	1.5369	-0.0131	0.81
	-0.6	-0.6042	-0.0042	0.28	-0.6007	-0.0007	0.05
	0.8	0.8096	0.0096	0.49	0.8178	0.0178	0.89
	0.3	0.2856	-0.0144	0.73	0.2914	-0.0086	0.43
F	1.031	0.9837	-0.0473	1.74	0.9845	-0.0465	1.76
	-0.630	-0.5879	0.0421	1.31	-0.5846	0.0454	1.48
	0.524	0.5008	-0.0232	1.58	0.5061	-0.0179	1.28
	-0.375	-0.5155	-0.1405	3.00	-0.5157	-0.1407	2.90
	0.8	0.8461	0.0461	1.37	0.8483	0.0483	1.44
	0.3	0.3214	0.0214	0.65	0.3185	0.0185	0.57
T	0.2	0.2112	0.0112	0.83	0.2106	0.0106	0.75
	0.4	0.3867	-0.0133	1.04	0.3919	-0.0081	0.60

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
E	AR	0.1116	1.00	0.1119	0.1144	1.02	0.1151
	AR	0.1060	0.96	0.1060	0.1095	0.99	0.1095
	MA	0.1372	1.05	0.1375	0.1406	1.08	0.1417
	MA	0.1395	1.13	0.1402	0.1428	1.16	0.1431
F	AR	0.1925	1.09	0.1982	0.1869	1.06	0.1926
	AR	0.2276	1.07	0.2315	0.2173	1.03	0.2220
	AR	0.1039	1.03	0.1064	0.0991	0.98	0.1007
	EV	0.3310	1.28	0.3596	0.3433	1.33	0.3710
	MA	0.2383	1.21	0.2427	0.2367	1.20	0.2416
	MA	0.2320	1.35	0.2330	0.2312	1.35	0.2319
T	MA	0.0954	0.99	0.0960	0.1002	1.04	0.1007
	MA	0.0907	1.00	0.0916	0.0963	1.06	0.0966

The tables reveal that the differences between the estimators are not very great in any sample. Conclusions on the significance of biases are similar for both estimators and a t-test performed on the equality of means reveals no case in which the null hypothesis is rejected at a 10% level. For variances and RMSE the ranking is not clear. B-J has a smaller variance and RMSE for Models E and T in all sample sizes but the ranking based on Model F fluctuates, with PH having smaller variance for all parameters when N = 40 and all except γ_0 when N = 100. For N = 70 B-J dominates. A test for equality of variances reveals that the F value is not significant at the 5% level for any experiment but that values around 1.46 were obtained for Model E when N = 40. The evidence from the tables leads to the conclusion that the inclusion of e* in the objective function will not affect the distributions greatly, and what is more important, will not provide an escape from the quandry raised by Model T.

4.6 A Comparison of A.F.D. and PH

Three models are selected for the comparison of this section - Models A, T and U below.

MODEL U

$$y(t) = (1 + 0.5L)\, e(t).$$

The basis of choice resides in the trade-off between the number of replications and parameters if a fixed computational burden is assumed. As the number of function evaluations required to form the Hessian was $\frac{5k(k+1)}{2}$ (where k is the number of parameters) and each evaluation computed $N/2$ cosine terms the computational cost of A.F.D. could become

heavy for quite small k. It was felt that the better alternative was to perform 50 replications rather than adopt more realistic models although Model T was selected mainly to investigate whether A.F.D. would display the peculiar behaviour of the time domain estimators. The results of all experiments are reported in Tables 4.19, 4.20 and 4.21.

Table 4.19

Summary Statistics for A.F.D. and PH

Models A, T and U : Sample Size N = 40

PART A

Model	P	A.F.D.			PH		
		M	B	B/SE	M	B	B/SE
A	0.8	0.7223	-0.0777	5.01	0.741	-0.059	3.28
	0.5	0.4437	-0.0563	2.48	0.550	0.050	2.07
T	0.2	0.1908	-0.0092	0.38	0.1869	-0.0131	0.53
	0.4	0.3750	-0.0250	1.14	0.5261	0.1261	3.45
U	0.5	0.4857	-0.0143	0.61	0.5208	-0.0208	0.85

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
A	AR	0.1094	1.04	0.1341	0.1264	1.21	0.1395
	MA	0.1602	1.06	0.1698	0.1692	1.12	0.1764
T	MA	0.1691	1.11	0.1693	0.1750	1.15	0.1755
	MA	0.1548	1.08	0.1568	0.2582	1.80	0.2873
U	MA	0.1651	1.21	0.1657	0.1725	1.26	0.1736

Table 4.20

Summary Statistics for A.F.D. and PH

Models A, T and U : Sample Size N = 70

PART A

Model	P	A.F.D.			PH		
		M	B	B/SE	M	B	B/SE
A	0.8	0.7348	-0.0652	4.76	0.759	-0.041	3.18
	0.5	0.5291	0.0291	1.43	0.544	0.044	2.60
T	0.2	0.2174	0.0174	1.07	0.2180	0.0180	1.06
	0.4	0.3496	-0.0504	2.30	0.4178	0.0178	0.90
U	0.5	0.5018	0.0018	0.12	0.5280	-0.0280	1.84

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
A	AR	0.0969	1.24	0.1176	0.0905	1.15	0.0993
	MA	0.1433	1.27	0.1462	0.1182	1.05	0.1261
T	MA	0.1159	1.01	0.1170	0.1204	1.04	0.1217
	MA	0.1034	0.96	0.1150	0.1399	1.29	0.1410
U	MA	0.1172	1.13	0.1172	0.1076	1.04	0.1112

Table 4.21

Summary Statistics for A.F.D. and PH

Models A, T and U : Sample Size N = 100

PART A

Model	P	A.F.D.			PH		
		M	B	B/SE	M	B	B/SE
A	0.8	0.7476	-0.0524	4.56	0.762	-0.038	3.49
	0.5	0.4945	-0.0055	0.35	0.541	0.041	3.06
T	0.2	0.2069	0.0069	0.51	0.2106	0.0106	0.75
	0.4	0.3557	-0.0443	3.85	0.3919	-0.0081	0.60
U	0.5	0.5114	0.0114	0.95	0.5240	-0.0240	1.88

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
A	AR	0.0817	1.25	0.0970	0.0763	1.17	0.0852
	MA	0.1100	1.17	0.1101	0.0942	1.00	0.1027
T	MA	0.0962	1.00	0.0964	0.1002	1.04	0.1007
	MA	0.0810	0.90	0.0923	0.0963	1.06	0.0966
U	MA	0.0854	0.99	0.0861	0.0907	1.05	0.0938

A.F.D. has variable sampling performance. For Models A and U A.F.D. compares favourably with PH in a R.M.S.E. sense but, somewhat surprisingly, the bias on $\hat{\alpha}_4$ in Model T seems to rise with sample size, becoming very significant when N = 100. This is rather interesting as the same inexplicable feature was noted for PH/O and, when linked with the knowledge that there was no evidence of the boundary condition being met for $\hat{\alpha}_4$ in any trial with A.F.D., creates a suspicion that the two estimators may be very similar in their sampling performance. Certainly both converged to the answers summarized in their respective tables from starting values in the vicinity of 0.9 - 0.95.

Computationally A.F.D. is very inferior to PH. Table 4.22 illustrates this with the C.P.U. time for 50 replications of Models T and U.

Table 4.22

Computation Time for A.F.D. and PH (Mins.)

Models T and U : All Sample Sizes

Model	Sample	A.F.D.	PH
T	40	7.58	5.94
	70	13.96	7.05
	100	22.47	8.05
U	40	2.85	0.86
	70	6.16	1.30
	100	11.08	1.79

For small and moderate sized samples the computational advantage of PH is substantial, but as the sample sizes grows A.F.D. begins to narrow the gap. The reason for this is twofold. Firstly the periodogram may be constructed by Fast Fourier Transform techniques so that the order of computation is N log N rather than N^2, and secondly it becomes possible to replace the periodogram by the spectrum so that the number of ordinates over which summation occurs may be very much less than N. For monthly series in which the number of observations may be 200-300 there are likely to be less objections to the use of A.F.D. on computational grounds, but the restriction to quarterly data imposed on the thesis precluded any testing of this.

4.7 Some Parametric Models of Economic Time Series

This section is devoted to a comparative analysis of the parameter estimates from each estimator when applied to models of two economic time series. The experiment was designed with two objectives in mind.

(a) To illustrate the application of the algorithms discussed in earlier sections to actual time series rather than to synthetic data.

(b) To examine the performance of the estimators in the light of the Monte Carlo studies.

The two series chosen were:

(1) Quarterly Personal Consumption Expenditure on Food ($m) from September 1950 to June 1970 inclusive.[3] There are eighty observations in the series which will be referred to hereafter as Food.

(2) Quarterly New Money Raisings by Listed Companies from June 1954 to June 1969 inclusive.[3] There are sixty observations in the series which will be designated as New Money.

The first of the two series was chosen to reflect the consumption bias of the thesis and it is re-estimated as part of a system of equations in the following chapter, while the second series provides a striking example of the difficulties arising from the presence of unit roots in the M.A. A number of reasons for expecting such a root in this model will be advanced in Chapter 6.

Some attention was paid to the specification of a suitable parametric model for Food and the stages leading to a final choice are outlined in Section 5.5. Suffice it to say that the model selected was

$$(1-L)\,(1-L^4)\,y(t) = (1+\alpha_4 L^4)\,e(t) \qquad (4.1).$$

Table 4.23 incorporates the estimates of α_4, the sample standard

---

[3] Supplied by the Commonwealth Statistician, Bureau of Census and Statistics, Canberra, Australia and listed in Appendix 11.

deviations and the residual variances from employing each estimator. Also included is the asymptotic standard deviation of the M.L. estimator evaluated by substituting $\hat{\alpha}_4$ into the formula for this.

Table 4.23

Equation (4.1) fitted to Food

All Estimators

Estimator	$\hat{\alpha}_4$	$\delta_{\hat{\alpha}_4}$	ASD	$\hat{\sigma}^2$
PH/O	-0.3647	0.1198	0.1080	94.18
PH	-0.4363	0.1216	0.1039	87.60
B-J	-0.4239	0.1215	0.1046	97.70
A.F.D.	-0.4702	0.1053	0.1019	87.78
A.T.D.	-0.4279	0.1283	0.1044	92.42

An assessment of Table 4.23 reveals only slight differences between the estimators as evidenced both by the magnitude of the parameters and the 95% confidence intervals which, when constructed around each estimate, encompass all other estimates. The most unexpected feature of the table is the considerable variability in $\hat{\sigma}^2$, with the discrepency between the B-J and PH values being most surprising but the general impression is that all estimators behave in a similar fashion to the Monte Carlo experiments. Finally it is encouraging to note the excellent agreement between sample and asymptotic standard deviations for all estimators.

New Money was not prespecified in any rigorous manner but the equation fitted was that given in Brewer et al [14].

$$(1-L)\ (1-L^4)\ y(t) = (1 + \alpha_1 L)\ (1 + \alpha_4 L^4)\ e(t) \qquad (4.2)$$

These authors fitted (4.2) to the logarithm of the New Money series in the context of a seasonal adjustment of this data. Table 4.24 presents the estimates of α_1 and α_4, their standard deviations and the residual variances obtained from each estimator.

Table 4.24

Equation (4.2) fitted to New Money

All Estimators

Estimator	$\hat{\alpha}_1$	$\sigma_{\hat{\alpha}_1}$	$\hat{\alpha}_4$	$\sigma_{\hat{\alpha}_4}$	$\hat{\sigma}^2$
PH/O	-0.1436	0.1417	-0.5967	0.1186	635.1
PH	-0.2195	0.1481	-0.9998	0.1187	435.9
B-J	-0.1585	0.1491	-0.9994	0.1392	528.7
A.F.D.	-0.2213	0.1543	-0.6124	0.0998	583.0
A.T.D.	-0.1805	*	-0.9999	*	480.8

What is of greatest import in Table 4.24 is that the time domain estimators (excluding PH/O) act as if there were a unit root in the M.A., and in seeking to break the boundary condition duplicate the behaviour observed with Model T when N = 40. Also familiar from the Model T experiments are the absence of a unit root in the M.A. transfer function when estimated by A.F.D. and PH/O and convergence to the values of Table 4.24 from a wide range of initial points for all estimators. However there is a conceptual difference between the models. For Model T it was known that the true transfer function did not have unit roots, whereas (as Chapter 6 will argue) there are some grounds for believing that (4.2) is a misspecification and it is this that is inducing the unit root. If so, the failure of A.F.D. and PH/O to indicate it must

create some doubts about their utility (or at least robustness against misspecification).

Finally a comment on the standard deviations associated with A.T.D.. As $\hat{\alpha}_4 \to 1$ it was not possible to invert Ω owing to the presence of a singular matrix, and the asterisks signify this failure. To understand the causes of the singularity reference should be made to the theorem of Section 1.5.1 that identifies the maximum and minimum eigenvalues of Ω with the spectral ordinates $f_{uu}(\lambda)_{MAX}$ and $f_{uu}(\lambda)_{MIN}$. For a M.A. process of the form $u(t) = (1+\alpha_1 L)(1+\alpha_4 L^4)\, e(t)$ it is easily seen that $\lim_{\alpha_4 \to 1} (f_{uu}(\lambda)_{MAX} / f_{uu}(\lambda)_{MIN}) = \infty$, so that the condition number of the matrix Ω becomes extremely large.

4.8 Conclusion

Chapter 4 has compared a number of estimators to that proposed by Phillips. Generally the impression has been that this estimator is at least no worse than others proposed in the literature and its computational efficiency leads to its acceptance for use in empirical investigation. Perhaps the major qualification to such a statement is that PH/O is likely to be valuable in situations when it is desirable to limit the number of parameters that must be estimated from the data, and, as a partial check on the validity of applying PH/O is available from an inspection of the estimated M.A. root, will frequently be used in later chapters. Certainly in attempting to decide on a reasonable specification PH/O is the fastest and cheapest algorithm to use.

CHAPTER 5 The Estimation of Non-Simultaneous Systems of ARMAX Equations

5.1 Introduction

Previous chapters have focussed on the development of various single-equation estimators and the evaluation of their sampling distributions under a variety of conditions. However there are obvious deficiencies in the construction and use of single equation estimators. Two that have assumed some importance in the econometric literature are:-

(i) The possibility of a feedback relationship between variables in a system of equations causing a member (or members) of the regressor set to be correlated with the disturbances. Haavelmo [37] studied the properties of O.L.S. in this situation and, in showing that it was not consistent, established a "simultaneous equation bias".

(ii) If substantial intercorrelation exists between time series, and it is believed to be "constant", efficient predictors and estimators must incorporate this feature. Zellner [123] was the first to analyse this and under the heading of "seemingly unrelated equations" furnished an efficient estimator.

This chapter is concerned with (ii). Although the solution of (ii) may be regarded as a necessary prelude to the solution of (i) the presence of simultaneous relationships raises a number of issues that are beyond the scope of the thesis.

Having defined the area of study a short outline of the chapter will now be presented. The following section deals with a generalization of the Phillips estimator of single equations to a system of equations and there will be some discussion concerning the program written to computerize this. Section 5.3 follows the dichotomy of previous

chapters in deriving an analogous estimator in the frequency domain, and obtaining as an extension of this the asymptotic covariance matrix of the efficient estimator which will be used, as in Chapters 3 and 4, to assess the Monte Carlo experiment of Section 5.4. Lastly, owing to the considerable computational cost incurred in Monte Carlo experimentation, both estimators are applied to a system of consumption equations thereby enabling a comparison of the sampling and asymptotic variances.

5.2 A Generalized Phillips Estimator

5.2.1 Derivation

As mentioned the estimator to be proposed belongs to the class of "seemingly unrelated estimators" introduced by Zellner. Zellner himself has generalized the methodology to take account of autocorrelation in the disturbances [124], and recently Kmenta and Gilbert [62] have conducted a number of Monte Carlo experiments upon the estimator with models in which the disturbance process follows a first order autoregression. This study revealed that the gains in efficiency to be derived from a recognition of the autocorrelation structure in the residuals was substantial, and that even in small samples (ten) these gains were such that imposition of the non-linear restrictions implied by the disturbance format would always be wise. In the same spirit the extension to be suggested in this section prescribes the disturbance term as a moving average, and an iterative algorithm is described that will yield M.L. estimates of all parameters.

Some notation is required. All arrays etc. will bear the same meaning as they possessed in the section on the Phillips estimator in Chapter 2 except that if subscripted, reference is to the expression

appropriate to a single equation, and if not, to the corresponding system matrix. As an example let the j'th equation of a G equation system be ARMAX (p,r,q) and this will be written in lag operator form as [1]

$$B_j(L)y_j(t) = C_j(L)x_j(t) + A_j(L)e_j(t) \qquad (5.1),$$

corresponding to equation (2.3).

The system of G ARMAX equations will be

$$\begin{aligned} B_1(L)y_1(t) &= C_1(L)x_1(t) + A_1(L)e_1(t) \qquad (5.2),\\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots\\ B_G(L)y_G(t) &= C_G(L)x_G(t) = A_G(L)e_G(t) \end{aligned}$$

or using the rule relating to subscripts we have the system form

$$B(L)\, y(t) = C(L)\, x(t) + A(L)\, e(t) \qquad (5.3),$$

where $y(t)$ is a (Gxl) vector $\{y_1(t) \ldots y_G(t)\}$

$x(t)$ is a (Gxl) vector $\{x_1(t) \ldots x_G(t)\}$,

$e(t)$ is a (Gxl) vector $\{e_1(t) \ldots e_G(t)\}$.

Three restrictions will be imposed on (5.3).

(i) It is assumed that B(L) and A(L) are diagonal matrices. This is primarily designed to enable the application of the identification conditions of Chapter 1 to each equation in turn, but at the same time it is the most likely form to be encountered.

(ii) e(t) will be assumed normally distributed with

$$E(e_j(t)) = 0.$$

---

[1] As before the derivation will be for a single exogenous variable but this assumption is for convenience only.

$$E(e_i(t)\ e_j(t-s)) = \sigma_{ij} \qquad s = 0 \qquad i,j = 1,\ldots,G$$
$$= 0 \qquad s \neq 0$$

(iii) Finally there are P,R and Q - respectively the number of A.R., E.V. and M.A. - parameters in the system with N observations remaining after the lags in all variables are accounted for.

Equation (5.1) resembles (2.3) and by following the transformation to the matrix form set out in (2.12) (p 52) (5.2) may be rewritten as

$$\begin{aligned} y_1 &= Y_1\beta_1 + X_1\gamma_1 + M_1 e_1 + M_1^* e_1^* \\ &\ \ \vdots \\ y_G &= Y_G\beta_G + X_G\gamma_G + M_G e_G + M_G^* e_G^* \end{aligned} \qquad (5.4).$$

As a guide to the interpretation of (5.4) suppose that the first equation is ARMAX (p,r,q). Then β_1 will be (pxl) vector, γ_1 a (rxl) vector and e_1^* a (qxl) vector. The elements of each will be distinguished by a second subscript e.g. β_{12} is the second A.R. parameter in the first equation. However this fine division will only be required at the termination of Section 5.3 and in Appendix 9.

The system of equations may be arranged as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_G \end{bmatrix} = \begin{bmatrix} Y_1 & & 0 \\ & \ddots & \\ 0 & & Y_G \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_G \end{bmatrix} + \begin{bmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_G \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_G \end{bmatrix} + \begin{bmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_G \end{bmatrix} \begin{bmatrix} e_1 \\ \vdots \\ e_G \end{bmatrix}$$

$$+ \begin{bmatrix} M_1^* & & 0 \\ & \ddots & \\ 0 & & M_G^* \end{bmatrix} \begin{bmatrix} e_1^* \\ \vdots \\ e_G^* \end{bmatrix} \qquad (5.5),$$

$$\text{or } y = Y\beta + X\gamma + Me + M^*e^* \qquad (5.6).$$

Under the previous assumption that e is normally distributed with covariance matrix

$$E(ee^T) = \Omega \otimes I$$

where $\otimes$ is the Kronecker product, the likelihood function is

$$\log L\,(y/x, e^*, y^*, x^*;\ \Omega,\theta) = \text{const.} + \log\det\,(\Omega^{-\frac{1}{2}} \otimes I) - \frac{1}{2} e^T(\Omega^{-1} \otimes I)e \qquad (5.7),$$

where x^*, e^* and y^* are the pre-period values of x,e and y.

Because $\log\det\,(\Omega^{-\frac{1}{2}} \otimes I)$ is of lower order than the exponent, asymptotically it is permissible to concentrate upon

$$S = e^T(\Omega^{-1} \otimes I)e \qquad (5.8).$$

Grouping the parameter vectors β, γ, α and e^* into the vector θ and denoting the matrix of first derivatives $\frac{\partial e}{\partial \theta}$ by Z, the Gauss-Newton algorithm yields the following sequence of iterations for minimizing (5.8) with respect to θ.

$$\theta^{(n)} - \theta^{(n-1)} = -(Z^T(\Omega^{-1} \otimes I)Z)^{-1} Z^T(\Omega^{-1} \otimes I)e \qquad (5.9).$$

The derivative matrix Z has the following elements.

$$\frac{\partial e_k}{\partial \beta_k} = -M_k^{-1} Y_k \qquad k = 1,\ldots, G$$

$$\frac{\partial e_k}{\partial \gamma_k} = -M_k^{-1} X_k \qquad k = 1,\ldots, G \qquad (5.10)$$

$$\frac{\partial e_k}{\partial \alpha_k} = -M_k^{-1} \bar{E}_k \qquad k = 1,\ldots, G$$

$$\frac{\partial e_k}{\partial e_k^*} = -M_k^{-1} M_k^* \qquad k = 1,\ldots, G$$

where $\bar{E} = \begin{bmatrix} \bar{E}_1 & & 0 \\ & \ddots & \\ 0 & & \bar{E}_G \end{bmatrix}$ and $\bar{E}_j$ $(j = 1,\ldots,G)$ is the matrix with elements described on page 53 of Section 2.3.1.

The Gauss-Newton algorithm is iterated until $e^T(\Omega^{-1} \otimes I)e$ changes by less than a prespecified amount (normally 10^{-4} of its previous value). At the minimum an estimate of the covariance matrix of the M.L. estimator $V_{\hat{\theta}\hat{\theta}}$ is

$$V_{\hat{\theta}\hat{\theta}} = (Z^T(\Omega^{-1} \otimes I)Z)^{-1} \qquad (5.11).$$

5.2.2 Computation

There are a number of topics to be discussed under this heading.

(i) The Covariance Matrix (Ω)

Examining (5.9) it is clear that the algorithm is not operational unless an estimate of Ω is available. It might be possible to include the elements of Ω in the parameter set and iterate upon these as well, but the properties of the resulting estimates are not well defined. In lieu of this we shall adopt a version of the solution given originally by Zellner.

Consider the two strategies:

(a) Calculate $\hat{\theta}^*$ from (5.9) with a consistent estimate (Ω^*) of Ω inserted.

(b) Calculate $\hat{\theta}$ from (5.9) with the true value of Ω inserted.

Then Zellner has shown [123 p 353] that $\hat{\theta}^*$ is asymptotically efficient relative to $\hat{\theta}$ if the regressors are uncorrelated with e. As the regressors contain only predetermined variables this condition will be satisfied, allowing the residuals from applying the single equation ARMAX

estimator to each equation in turn to be used in constructing an estimate of Ω. Estimates of σ_{ij} were then obtained by the following formula.[2]

$$\hat{\sigma}_{ij} = \sum_{t=0}^{N-1} \hat{e}_i(t)\, \hat{e}_j(t)/N \qquad i,j = 1,\ldots,G.$$

(ii) Estimates of e*

One of the disconcerting features of the Phillips estimator is the need to make some estimate of the lagged disturbance vector e*. If this is done by the generalized Phillips method, an inversion of order at least (P + 2Q + R) is required at each iteration. In practice however the order of the inversion may be much larger than this because the number of M.A. parameters may be less than the order of the M.A. process i.e. some of the intermediate M.A. parameters are zero e.g. $(1 + \alpha_1 L)(1 + \alpha_4 L^4)$. As the determinant of the order of e* is the order of the M.A. polynomials, a very large number of e* parameters may be present even though the number of other parameters is small.

Two aspect of this must be mentioned. Firstly the computational cost may be very high when the number of parameters becomes large. Secondly rounding errors in the inversion routine will rise as the order of the inversion rises and may lead to incorrect estimates. Because the parameters e* are not of fundamental interest it was argued that they should not be allowed to prohibit the extraction of others leading to the following three solutions.

(a) Assume e* = 0. As suggested in Chapter 2 this may be a near optimal strategy if the roots of the M.A. are not close to unity and/or the sample size is large. Certainly when faced with the need to consider systems estimators any preliminary analyses should be carried out under

---

[2] Note that this is a consistent estimate and differs from the unbiassed estimate normally adopted. The choice reflects the conditions of Zellner's theorem.

this assumption.

(b) Another solution is to be had by assigning the consistent estimates of e^* resulting from an application of the single equation estimator to each equation in turn. The impact of this proposal is difficult to assess. If there was only slight variation in the parameter estimates in the transition from a single equation to a systems of equations context, then it is likely that the single equation estimates of e^* will be closer to the M.L. values than zero would be. In such a case (b) would be (almost) computationally equivalent to (a) and it would be preferred. Generally both (a) and (b) were tried in later sections and (a) was normally selected but obviously investigations should be made of the choice by Monte Carlo methods.

(c) Concentration of the likelihood with respect to e^* might be used profitably in a system context, but as was found in the single equation case it is important that there be low order M.A. processes in order to avoid the sharp rise in computation time which is a concomitant of the inversions associated with an evaluation of the S function. It is interesting to note that in a recent article Lawton and Sylvestre [64] have advocated the concentration of S with respect to all parameters that enter e linearly. From (5.10) it can be seen that only α enters in a non-linear fashion so that there is a good case for concentration with respect to β, γ and e^*. The authors refer to non-linear functions in which concentration has improved convergence and also claim a reduced computational burden, but as was seen in Chapter 2 this does not necessarily follow.

(iii) Numerical Derivations and Another Algorithm

For the Monte Carlo study described below the Gauss-Newton formula with derivatives given by (5.10) was employed. However when applying it to actual data it was easier to obtain flexibility by using numerical derivatives. Provided the Gauss-Newton algorithm was adopted, the computational cost was not excessive for any of the formulations examined in 5.5.

5.3 A Frequency Domain Estimator

The purpose of this section is twofold. Primarily we will be concerned with the provision of a generalized version of A.F.D., but a secondary purpose will be to derive the asymptotic covariance matrix of the M.L. estimator. Such a division closely parallels that of Section 2.4, but possession of the asymptotic covariance matrix is of heightened interest in this chapter in that it is efficiency gains that are of paramount importance. Only a system of ARMA equations will be considered in detail, but the generalization to ARMAX equations would proceed along the same lines.

To begin we adopt the principle espoused in the preceding section that all non-subscripted variables are of the system and all subscripted of one equation. Extending the nomenclature of Section 2.4 in this way the system of ARMA equations may be written as

$$\begin{aligned} B_1(\lambda)\, Y_1(\lambda) = u_1(\lambda) &= A_1(\lambda)\, \varepsilon_1(\lambda) \\ \vdots \qquad\qquad \vdots &\qquad \vdots \\ B_G(\lambda)\, Y_G(\lambda) = u_G(\lambda) &= A_G(\lambda)\, \varepsilon_G(\lambda) \end{aligned} \tag{5.12},$$

where $Y_j(\lambda)$, $\varepsilon_j(\lambda)$ are (Nx1), $B_j(\lambda)$ and $A_j(\lambda)$ are (NxN) matrices (j=1,..,G) where N is the number of frequency bands in the range $0 < \lambda \leq 2\pi$ and

all symbols are defined in (2.24) - (2.26). We note that $A_j(\lambda)$ and $B_j(\lambda)$ have the form

$$A_j(\lambda) = \begin{bmatrix} A_j(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & A_j(\lambda_N) \end{bmatrix} \quad B_j(\lambda) = \begin{bmatrix} B_j(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & B_j(\lambda_N) \end{bmatrix},$$

meaning that $A_j(\lambda_k)$ is in row k, column k of the matrix $A_j(\lambda)$. By writing (5.12) in full it may be verified that it is a system of NG equations with each equation composed of scalars alone e.g. the first would be

$$B_1(\lambda_1)\, Y_1(\lambda_1) = A_1(\lambda_1)\, \varepsilon_1(\lambda_1).$$

The system can be arranged as

$$\begin{bmatrix} B_1(\lambda) & & 0 \\ & \ddots & \\ 0 & & B_G(\lambda) \end{bmatrix} \begin{bmatrix} Y_1(\lambda) \\ \vdots \\ Y_G(\lambda) \end{bmatrix} = \begin{bmatrix} u_1(\lambda) \\ \vdots \\ u_G(\lambda) \end{bmatrix} = \begin{bmatrix} A_1(\lambda) & & 0 \\ & \ddots & \\ 0 & & A_G(\lambda) \end{bmatrix} \begin{bmatrix} \varepsilon_1(\lambda) \\ \vdots \\ \varepsilon_G(\lambda) \end{bmatrix} \quad (5.13),$$

or $B(\lambda)\, Y(\lambda) = u(\lambda) = A(\lambda)\, \varepsilon(\lambda)$ (5.14)

with $B(\lambda)$, $A(\lambda)$ as (NGxNG) matrices and $Y(\lambda)$, $\varepsilon(\lambda)$ as (NGx1) vectors.

Before proceeding further some definitions are required.

$I_{u_k u_\ell}(\lambda_j) = \frac{1}{2\pi N}\, u_k(\lambda_j)\, \bar{u}_\ell(\lambda_j)$ is the cross periodogram between u_k and u_ℓ at frequency λ_j ($j = 1,\ldots,N$).

$I_{u_k u_\ell}(\lambda)$ = an (NxN) diagonal matrix with $I_{u_k u_\ell}(\lambda_j)$ ($j = 1,\ldots,N$) as diagonal elements.

$I_{uu}(\lambda)$ = the system matrix (NGxNG) composed of all $I_{u_k u_\ell}(\lambda)$ matrices

$$\text{i.e. } I_{uu}(\lambda) = \begin{bmatrix} I_{u_1u_1}(\lambda) \cdots I_{u_1u_G}(\lambda) \\ \vdots \qquad\qquad \vdots \\ I_{u_Gu_1}(\lambda) \cdots I_{u_Gu_G}(\lambda) \end{bmatrix}.$$

$f_{uu}(\lambda)$ = the spectral density matrix (NG x NG) which is of the same form as $I_{uu}(\lambda)$ but is constructed from the cross spectral densities between disturbance terms.

Taking the properties of $\varepsilon(t)$ as those of Section 5.2 Fishman [26 p 174 eq. (4.33a)] proves that the M.L. estimator minimizes (up to a constant of proportionality $\frac{1}{2}$)

$$S = \sum_{j=1}^{N} \sum_{k,\ell=1}^{G} I_{u_k u_\ell}(\lambda_j) \, f^{\ell k}_{uu}(\lambda_j) \qquad (5.15).$$

Equation (5.15) may be written in system notation as

$$S = \mathrm{tr}(I_{uu}(\lambda) \, f^{-1}_{uu}(\lambda)) \qquad (5.16).$$

A number of standard relations may be used to simplify (5.16). One ((5.17a)) is definitional, and two ((5.17b) and (5.17c)) are to be found in texts on linear filtering theory.

$$|A(\lambda)|^2 = A(\lambda) \, \bar{A}(\lambda) \qquad (5.17a)$$

$$f_{uu}(\lambda) = \frac{1}{2\pi} \, A(\lambda) \, (\Omega \otimes I) \, \bar{A}(\lambda) \qquad (5.17b)$$

$$I_{uu}(\lambda) = A^T(\lambda) \, \varepsilon^T(\lambda) \, \bar{\varepsilon}(\lambda) \, \bar{A}(\lambda) \qquad (5.17c)$$

Inverting (5.17b), and substituting the resulting expression along with (5.17c) into (5.16), we obtain

$$S = 2\pi \, \mathrm{tr}[A^T(\lambda)\varepsilon^T(\lambda)\bar{\varepsilon}(\lambda) \, \bar{A}(\lambda) \, \bar{A}^{-1}(\lambda)(\Omega^{-1} \otimes I)A^{-1}(\lambda)] \qquad (5.18).$$

As the order of matrices in the trace may be rearranged (5.18) reduces to

$$S = 2\pi tr\ [\varepsilon^T(\lambda)\ \bar{\varepsilon}(\lambda)\ (\Omega^{-1} \otimes I)] \qquad (5.19),$$

or

$$S = 2\pi \varepsilon^T(\lambda)\ (\Omega^{-1} \otimes I)\ \bar{\varepsilon}(\lambda) \qquad (5.20).$$

An inspection of the structure of (5.20) and (2.8) reveals a similar form and leads to the conclusion that minimization of (5.20) with respect to all parameters will lead to an Aitken estimator, and justifies the claim that it is an extension of A.F.D. to systems of equations.

The Gauss-Newton algorithm was selected to minimize (5.20). By the definition of $\varepsilon(\lambda)$ it is easily seen (from 5.12) that for the r'th equation $\varepsilon_r(\lambda_j) = \dfrac{Y_r(\lambda_j)B_r(\lambda_j)}{A_r(\lambda_j)}$, and from this the required derivatives with respect to the d'th A.R. and M.A. parameters of the r'th equation may be constructed.[3]

$$\frac{\partial \varepsilon_r(\lambda_j)}{\partial \beta_{rd}} = \frac{Y_r(\lambda_j)\, e^{id\lambda_j}}{A_r(\lambda_j)} \qquad \begin{matrix} r = 1,\ldots,G \\ j = 1,\ldots,N \\ d = 1,\ldots,p \end{matrix} \qquad (5.21)$$

$$\frac{\partial \varepsilon_r(\lambda_j)}{\partial \alpha_{rd}} = -\frac{Y_r(\lambda_j)B_r(\lambda_j)e^{id\lambda_j}}{A_r^2(\lambda_j)} \qquad \begin{matrix} r = 1,\ldots,G \\ j = 1,\ldots,N \\ d = 1,\ldots,q \end{matrix} \qquad (5.22)$$

Appendix 9 shows that asymptotically $\sqrt{N}\begin{pmatrix}\hat{\beta}-\beta \\ \hat{\alpha}-\alpha\end{pmatrix}$ has covariance matrix V^{-1} where V has elements

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{bmatrix}.$$

---

[3] For expository purposes it is assumed that each equation is ARMA (p,q), but it should be understood that the values of p and q may vary between equations.

For any parameter type the covariance between the d'th parameter in equation r and the g'th parameter in equation s will yield the following expressions for V.

$$V_{11}(k,\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(d-g)\lambda}\sigma_{rs}\sigma^{rs}}{B_r(\lambda)\,\bar{B}_s(\lambda)}\,d\lambda \qquad \begin{matrix} k,\ell = 1,\ldots,P \\ d,g = 1,\ldots,p \\ r,s = 1,\ldots,G \end{matrix}$$

$$V_{12}(k,\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(d-g)\lambda}\sigma_{rs}\sigma^{rs}}{B_r(\lambda)\,\bar{A}_s(\lambda)}\,d\lambda \qquad \begin{matrix} k = 1,\ldots,P \\ \ell = 1,\ldots,Q \\ r,s = 1,\ldots,G \\ d = 1,\ldots,p \\ g = 1,\ldots,q \end{matrix} \qquad (5.23)$$

$$V_{22}(k,\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(d-g)\lambda}\sigma_{rs}\sigma^{rs}}{A_r(\lambda)\,\bar{A}_s(\lambda)}\,d\lambda \qquad \begin{matrix} k,\ell = 1,\ldots,Q \\ r,s = 1,\ldots,G \\ d,g = 1,\ldots,q \end{matrix}$$

Equation (5.23) states the elements of the matrix to be inverted in order to obtain the covariance matrix of the efficient estimator. The corresponding elements for the single equation estimator are to be found from (5.23) by setting $\sigma_{rs} = 0$ if $r \neq s$ so that Ω will be diagonal, $\sigma^{ss} = 1/\sigma_{ss}$, and all cross terms vanish from (5.23) reducing it to (2.42).

Knowledge of the population values of the A.R. and M.A. parameters and the correlation matrix Ω, enables the determination of the asymptotic efficiency gains from an application of a systems estimator in preference to its single equation counterpart. As well as this it is interesting to note that a rough guide to the anticipated efficiency gains can be extracted from (5.23) by a substitution of the single equation estimates of β, α and Ω. Later it will be seen that the computation time for any systems estimator is high, so that it is of importance to assess the magnitude of efficiency gains in order to

balance the benefits against the expenses of usage.

5.4 A Monte Carlo Study of the Estimator

Following the methodology of previous chapters, information on the small sample distributional properties of the generalized Phillips estimator was sought via Monte Carlo experiments based on a six-parameter two-equation model.[4]

$$y_1(t) = 0.8\, y_1(t-1) + e_1(t) + 0.5\, e_1(t-1)$$

$$y_2(t) = 1.3\, y_2(t-1) - 0.6\, y_2(t-2) + e_2(t) + 0.6\, e_2(t-1) + 0.3\, e_2(t-2) \qquad (5.24)$$

A multivariate normal distribution for the disturbances $e_1(t)$ and $e_2(t)$ was constructed following the principles outlined in Naylor, Balintfy, Burdick and Chu [77 p 98]. This procedure requires two series of zero mean and unit variance Gaussian numbers, and as usual these were generated by GAUSS. The covariance matrix Ω was set at

$$\Omega = \begin{bmatrix} 9 & 1 \\ 1 & 1 \end{bmatrix}.$$

As has been emphasized in an earlier chapter the second of the equations is unlikely to be familiar to economists working with unfiltered data, and to this extent it may not be a good model to extract conclusions about economic phenomena. However referring to the results of Chapter 3 we might expect that the estimator would perform slightly better with this model than with a more realistic one.

Two sample sizes - 40 and 70-were selected and 50 replications were performed for each experiment. As the C.P.U. time claimed by each

---

[4] If the lagged epsilon terms are counted there were nine parameters.

replication was approximately 1 min., and each trial was repeated twice- once to generate initial estimates with the single equation estimator (SINGLE) and once with the system estimator (SYSTEM) - total computation time for each experiment was around 100 min.[5] Because of this, and the allied difficulty of constructing realistic systems that had a small number of parameters, only two experiments were performed so that the estimators are examined under very limited conditions. The experiments are summarized in Table 5.1.

Table 5.1

Summary Statistics of SYSTEM for the Model of Eq. (5.24)

Sample Sizes N = 40, 70

N	P	M	B	SE	|B/SE|	SD	ASD	SD/ASD
40	0.8	0.760	0.040	0.0141	2.84	0.0994	0.0967	1.03
	1.3	1.277	-0.023	0.0308	0.75	0.2175	0.1792	1.21
	-0.6	-0.587	0.013	0.0306	0.43	0.2163	0.1658	1.30
	0.5	0.546	0.046	0.0341	1.35	0.2412	0.1398	1.73
	0.6	0.650	0.050	0.0477	1.05	0.3376	0.2060	1.64
	0.3	0.397	0.097	0.0482	2.01	0.3408	0.1878	1.81
70	0.8	0.768	-0.032	0.0139	2.30	0.0883	0.0731	1.21
	1.3	1.285	-0.015	0.0202	0.74	0.1280	0.1355	0.94
	-0.6	-0.598	0.002	0.0193	0.10	0.1219	0.1253	0.97
	0.5	0.524	0.024	0.0230	1.04	0.1455	0.1056	1.38
	0.6	0.629	0.029	0.0296	0.98	0.1871	0.1557	1.20
	0.3	0.328	0.028	0.0262	1.07	0.1656	0.1420	1.17

Table 5.1 reveals that the systems estimator behaves in a similar fashion to SINGLE; as evidenced by the good results obtained for the second equation and the rather poorer results for the first. Chapter 3 concluded that the first equation always had large biases when estimated by a time domain method, and it is of some interest to

---

5 The designation SINGLE and SYSTEM derives from the calling routines.

observe that the biases of SYSTEM are smaller (particularly for the M.A. parameter).[6] As the starting values for the A.R. parameters were 0.7, 1.5, -0.75 (approximately the O.L.S. estimates for the first replication), a number of trials were repeated with different starting values. In all cases the algorithm converged to the same sum of squares but the parameters differed in the third decimal place. Another feature of Table 5.1 is the suggestion that asymptotic covariance formulae may not be reliable guides to the small sample covariances in samples of size 40 - at least for the M.A. parameters. Although it is noticeable that the ratio falls sharply as the sample size increases and, with the exception of α_1 in the first equation, are not statistically different from the asymptotic values in the larger sample, the experience of Chapter 3 would imply that in the transition to more realistic models the sample size needed for asymptotic theorems to hold is likely to be larger than 70. Nevertheless such results are hardly disheartening.

In view of the paucity of simulations attempted the only alternative source of information on the small sample distribution of SYSTEM is contained in other studies. Reference has already been made to the Kmenta and Gilbert article but the absence of lagged dependent variables in their system reduces the utility of their experiment for our purposes. Perhaps the success of the non-linear estimator in being ranked as superior to all others considered gives some support for the use of a non-linear algorithm in maximizing the system likelihood.

---

6 For the comparison see Tables 3.2, 3.3 and 3.4 (p 117-119).

A more recent study by Nelson [78] throws some light on the efficiency of SYSTEM when there are no M.A. parameters. This author attempted to deal with models possessing a M.A. term but seems to have been unable to develop a workable algorithm and therefore was forced to choose a two equation system of first order A.R. processes. He found that the expected gain in efficiency from an application of SYSTEM was never totally, but always at least 90% realized (for samples of size 30), and tabulations of the empirical distribution of the test statistic $(\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$ indicates that

> "t - tests may be carried out on second stage estimates in reasonably good conscience" [78 p 28].

Finally, although the experiments are not reported in full, it appears that with samples of size 100 efficiency gains could be predicted accurately by asymptotic formulae and that the above test statistic is likely to be normally distributed.

All in all the evidence presented in Table 5.1 and that available in Nelson's report is favourable to the proposition that SYSTEM might be applied in quite small samples. A more confident assertion must await further sampling experimentation, but in the meantime it is possible to gain some insight into the properties of the estimator by modelling some actual time series. At the same time this will allow a comparison with the generalized version of A.F.D. as the computer time required for any examination of the sampling properties of the latter estimator would be enormous.

5.5 An Experiment with Some Economic Time Series

5.5.1 Specification

This section considers the employment of systems estimators in the modelling of intercorrelated economic time series. A system of demand equations was chosen for this purpose, and the thirteen categories of consumption distinguished in the Australian National Accounts were collapsed to form five new series. The relationship between the two sets is given in Table 5.2.

Table 5.2

Construction of the Consumption Series from National Accounts Data

Series No.	National Account Designation	New Designation
1	Food	Food
2	Clothing and Drapery	Clothing
3	Electrical Goods, Other Household Durables, Electricity, Gas	Household Appliances
4	Cigarettes and Tobacco, Alcoholic Drinks, Postal and Telephone Fares, Purchase of Motor Vehicles, Other Goods and Services	Miscellaneous
5	Rent	Rent

One might quibble with some of the groupings employed e.g. Cigarettes and Alcohol could be categorized with Food, but as the exercise is to illustrate the use of the systems estimator there would not seem to be much to gain from putting effort into developing a new classification. Certainly for forecasting purposes it would be desirable to utilize all thirteen categories. Only four of the five aggregated series were adopted owing to the difficulties that arise

with rent (alluded to in Section 3.5) and some appreciation of their time behaviour is available from the correlograms presented in Table 5.3.

Table 5.3

Correlograms (8 lags) of all Consumption Series

Series	Lag							
	1	2	3	4	5	6	7	8
1	0.94	0.89	0.84	0.81	0.76	0.73	0.68	0.66
2	0.64	0.86	0.57	0.84	0.51	0.71	0.45	0.70
3	0.89	0.83	0.80	0.82	0.73	0.68	0.65	0.67
4	0.94	0.90	0.85	0.81	0.76	0.73	0.68	0.65

Of the four series in Table 5.3 only the second exhibits any peculiarities in the correlogram. The explanation of the cyclical pattern observed there resides in the six monthly cycle in Clothing caused by the summer-winter dichotomy of the Australian climate. All other series possess correlograms that by their slowly - declining nature suggest some form of non-stationarity in the data, and for Series 1,3 and 4 at least a first differencing filter was required to produce stationarity. As a first step it is logical to try both first and fourth differencing for these series. Such a filter is obviously inappropriate for series 2, but although a large number of others were applied (involving L^2 terms mainly) there was little improvement over that resulting from first and fourth differencing. Table 5.4 presents the correlograms of the first and fourth differenced series with the statistic $\frac{\hat{\rho}_k}{\sqrt{N}}$ (where $\hat{\rho}_k$ is the estimate of the k'th sample correlation coefficient) enclosed in brackets underneath. This statistic is

approximately distributed as Student's t with (N-k) degrees of freedom.[7]

Table 5.4

Correlogram (8 lags) of the Differenced Consumption Series

Series	Lag							
	1	2	3	4	5	6	7	8
1	-0.09	0.15	0.01	-0.23	-0.05	-0.08	0.11	-0.07
	(0.81)	(1.28)	(0.08)	(2.0)	(0.43)	(0.71)	(0.92)	(0.61)
2	-0.23	0.20	-0.15	0.04	-0.31	0.11	-0.28	0.26
	(2.02)	(1.73)	(1.34)	(0.31)	(2.66)	(0.97)	(2.39)	(2.28)
3	0.15	0.05	-0.14	-0.14	-0.32	-0.26	0.00	0.18
	(1.34)	(0.41)	(1.23)	(1.23)	(2.74)	(2.23)	(0.02)	(1.57)
4	-0.10	-0.09	0.08	-0.39	-0.19	0.26	-0.08	-0.14
	(0.86)	(0.75)	(0.72)	(3.37)	(1.63)	(2.28)	(0.72)	(1.18)

From Table 5.4 none of the individual t values are low enough to suggest that any of the differenced series is white noise with the possible exception of the first. A "portmanteau" test for white noise over k lags of the correlogram is available from forming

$Q = N \sum_{j=1}^{k} \hat{\rho}_j^2$ and testing Q as a $\chi^2(k)$ variable. For the four series Q has values 8.21, 27.84, 19.99 and 22.97 which, when compared to the 1%,

---

7 This test statistic is deficient in a number of ways. Box and Jenkins [9 p 35] state that the correct test statistic that all $\hat{\rho}_k$ are zero after ρ_q is given by forming

$N^{\frac{1}{2}} \hat{\rho}_k (1 + 2 \sum_{j=1}^{q} \rho_j^2)$ and testing this as t(N-k).

As ρ_j is unknown this cannot be formed, but extrapolating from the small $\hat{\rho}_k$ in Table 5.4, the discrepancy between the correct test and that in the text will probably not be great. When testing for white noise q = 0 and the two statistics coincide.

5% and 10% levels of $\chi^2(8)$ (20.1, 15.5 and 13.4 respectively), leads to the conclusion that white noise would be only likely for the first.

Owing to the correlation existing between neighbouring auto-correlation coefficients it is of some interest to examine the spectra of the differenced series. Figs 5.1, 5.2, 5.3 and 5.4 contain these for the four series. Parzen weights and twelve lags were used to construct all spectra and the resolution was not greatly improved by the employment of more lags than 12.

Theoretically the spectrum of a fourth order M.A. with α_4 being negative would have equal peaks at $\pi/4$ and $3\pi/4$ and a trough at $\pi/2$. Figure 5.1 is quite close to this pattern indicating that there may well be a fourth order M.A. present in the data (and thereby corroborating the significant value of $\hat{\rho}_4$ found), although the construction of confidence bands around the spectrum might invalidate this conclusion. None of the other spectra are compatible with such a simple hypothesis and this attests to the significant correlations at lags such as 5 and 6 in Series 3. In particular differencing has accentuated the peak at π that would be present in the original series as a concomitant of a six monthly cycle. This peak was even greater in the periodogram of the differenced series and created difficulties for A.F.D.,so that it was decided to eliminate Clothing as well.[8] There is little doubt that an adequate specification would require a great deal of analysis.

---

[8] Actually the obstacles were two in number. Firstly the peak at π in the periodogram, and secondly the near unit roots in the M.A. transfer function when various models were fitted. Inevitably the contribution to the sum of squares from π was in the vicinity of 10^{12} while near unit roots in $|A(\lambda)|^2$ accentuated rounding error to such an extent that, even when summed from $-\pi$ to π, the complex part was non-zero.

FIGURE 5.1

SPECTRUM OF FOOD AFTER FIRST

AND FOURTH DIFFERENCING

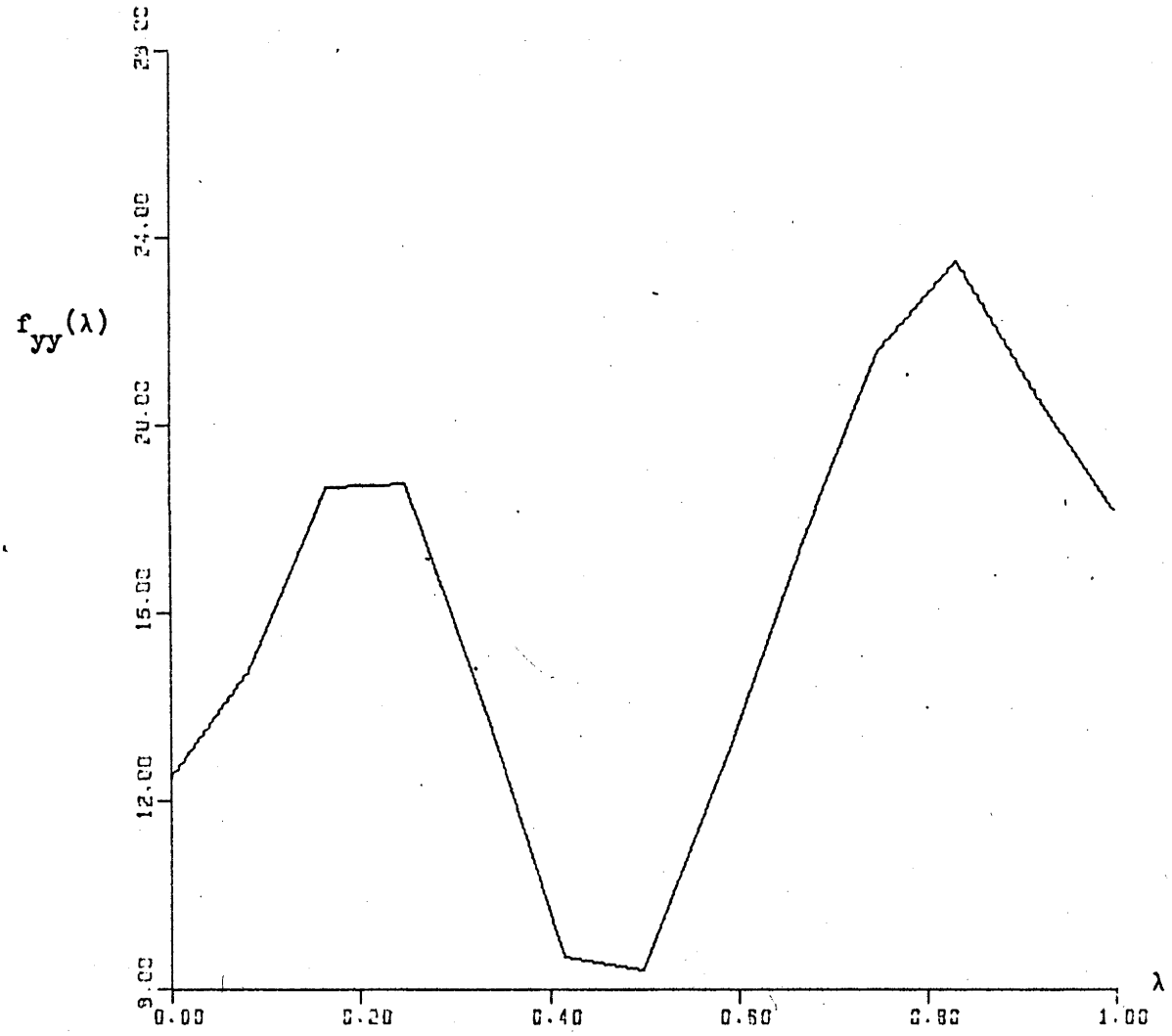


FIGURE 5.2

SPECTRUM OF CLOTHING AFTER FIRST

AND FOURTH DIFFERENCING

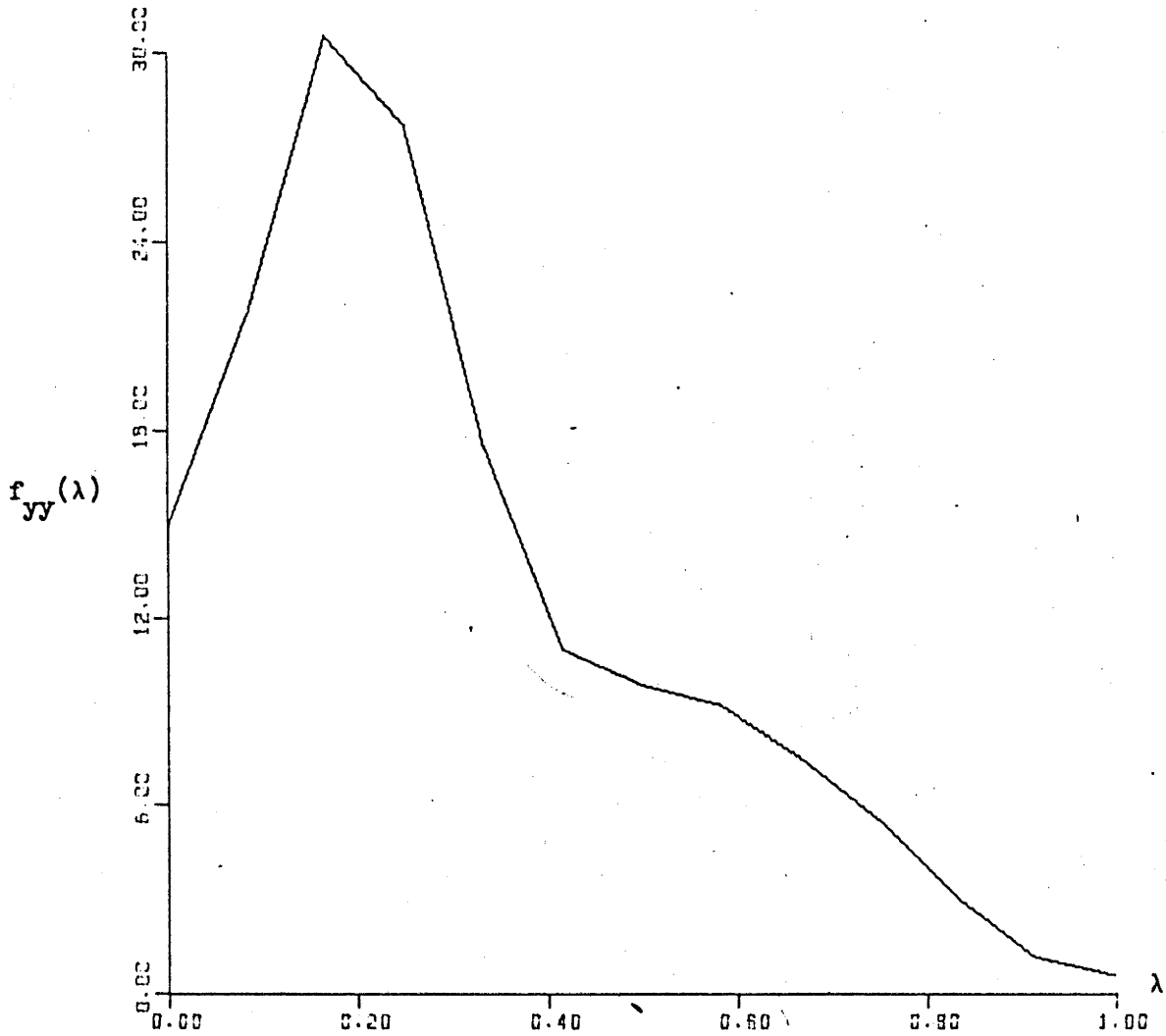


FIGURE 5.3

SPECTRUM OF HOUSEHOLD APPLIANCES AND

EQUIPMENT AFTER FIRST AND FOURTH DIFFERENCING

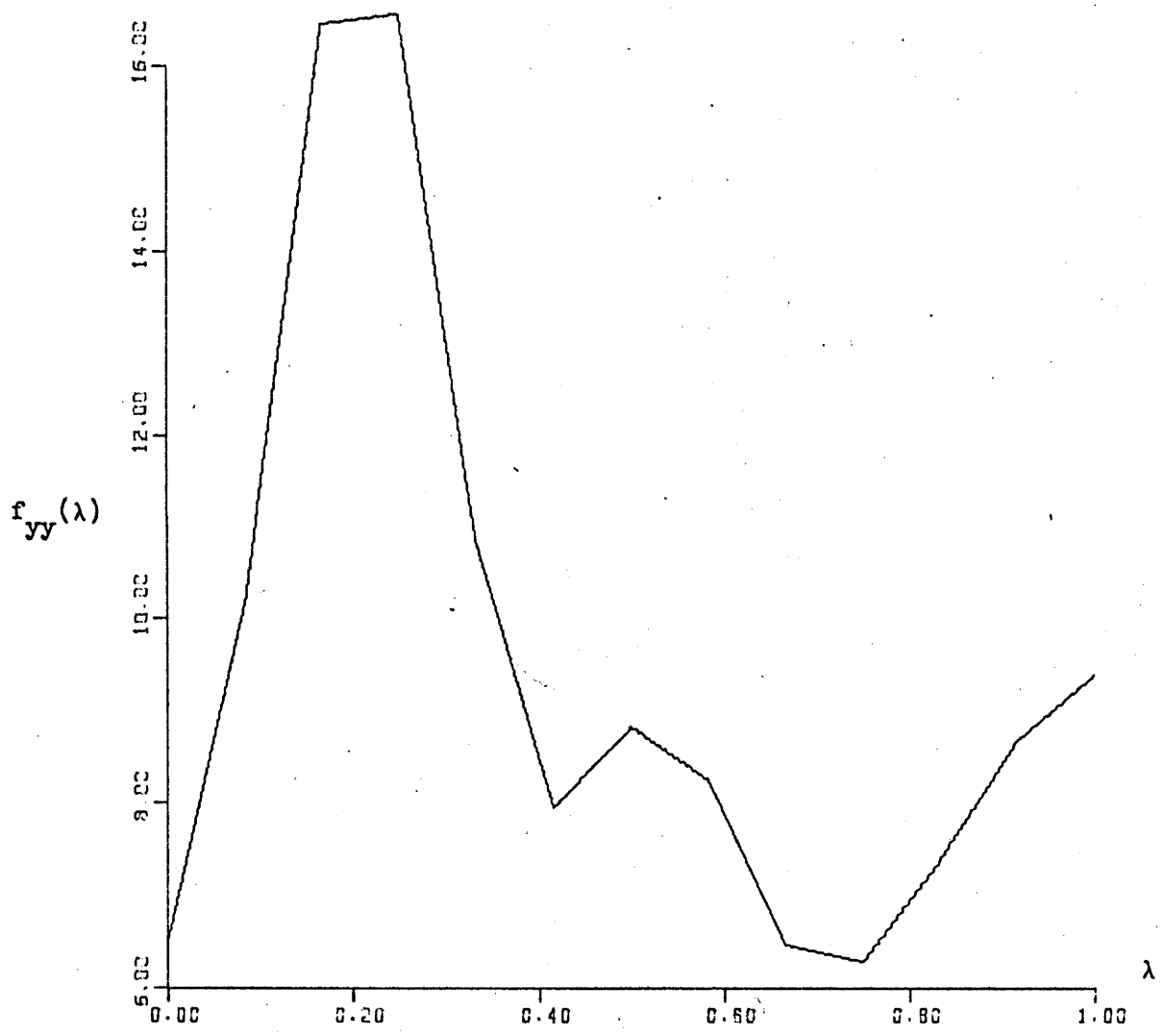
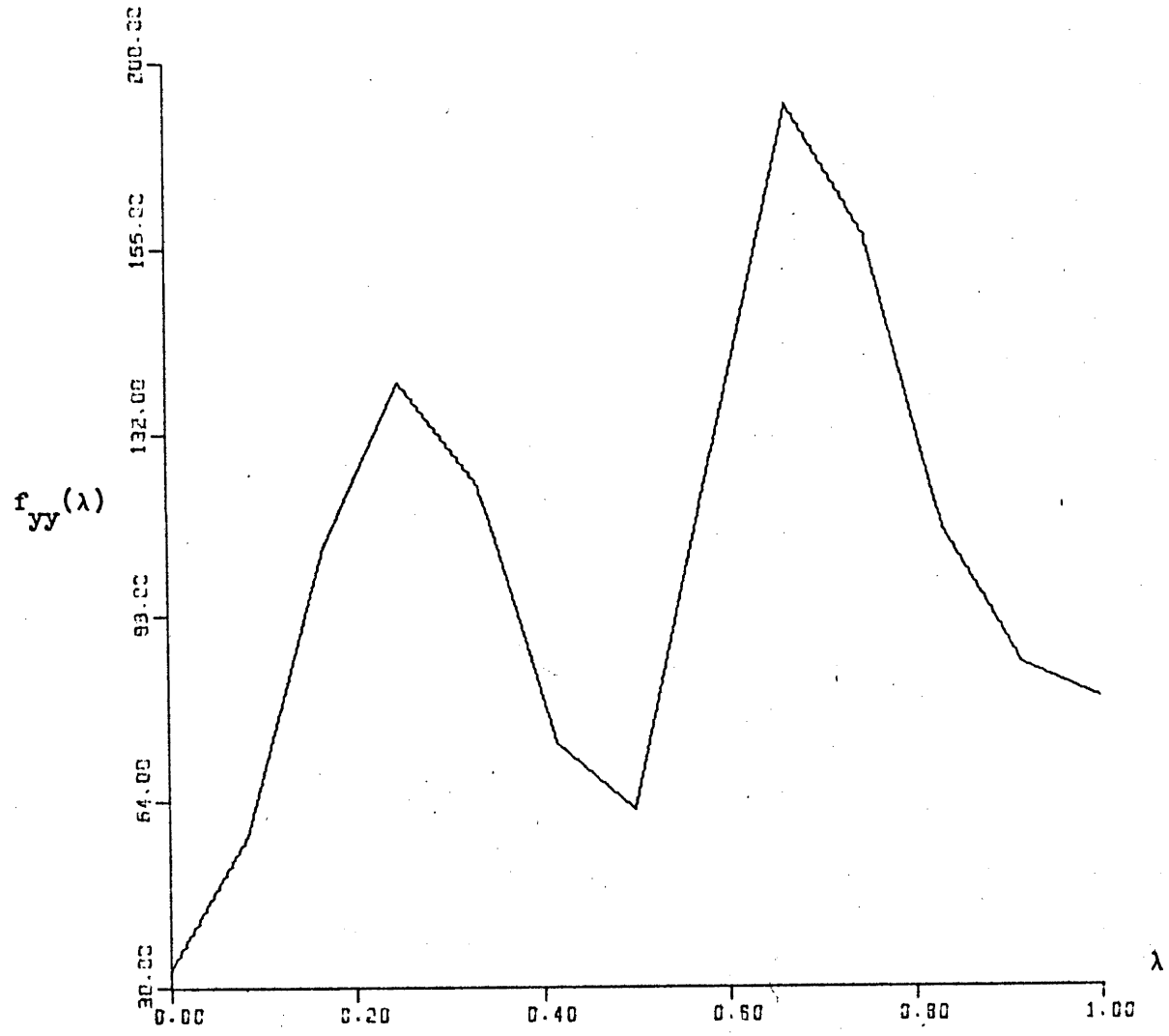


FIGURE 5.4

SPECTRUM OF MISCELLANEOUS AFTER FIRST

AND FOURTH DIFFERENCING



A variety of models might be fitted. From a consideration of Table 5.4 it was decided to fit the following to all series.

$$(1-L)(1-L^4)\, y(t) = \mu + (1+\alpha_1 L + \alpha_4 L^4 + \alpha_5 L^5 + \alpha_6 L^6)\, e(t) \quad (5.25)$$

One would expect that any inadequacy may be measured in the diagnostics. Table 5.5 gives the parameters, the t statistics, the residual variance and R^2 for Series 1,3 and 4 when estimated by SINGLE.

Table 5.5

Estimated Parameters of Equation (5.25)

Series 1,3 and 4

Series	$\hat{\mu}$	$\hat{\alpha}_1$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\sigma}^2$	R^2
Food	0.0701	-0.1755	-0.4062	-0.1460	-0.2709	89.97	0.996
	(0.22)	(1.43)	(3.34)	(1.15)	(2.17)		
Durables	0.1121	0.0945	-0.2748	-0.4449	-0.3176	44.16	0.992
	(0.49)	(0.73)	(2.18)	(4.05)	(2.65)		
Miscell-	1.2645	-0.3245	-0.5192	-0.3825	0.4125	493.78	0.996
aneous	(1.95)	(2.75)	(5.03)	(3.23)	(3.55)		

Table 5.5 shows that the mean is insignificant for both Food and Clothing but probably should be retained for Miscellaneous. Computing the roots of the M.A. polynomial it is striking that at least one of the roots is close to the boundary in all cases. To some extent this may be a consequence of a misspecification of the M.A. and therefore would be removed by a re-estimation of each equation. To decide which parameters should be retained the Wald test for removing sets of parameters was performed i.e. $\hat{\theta}^T V^{-1} \hat{\theta}$ (where V is the covariance matrix of $\hat{\theta}$) is distributed as $\chi^2(k)$ under the null hypothesis that k of the elements of $\hat{\theta}$ are zero. This process yielded the following equations:

1. $(1-L)(1-L^4)\ y(t) = (1+\alpha_4 L^4)\ e(t)$ (5.26a)

3. $(1-L)(1-L^4)\ y(t) = (1+\alpha_5 L^5 + \alpha_6 L^6)\ e(t)$ (5.26b)

4. $(1-L)(1-L^4)\ y(t) = \mu + (1+\alpha_1 L + \alpha_4 L^4 + \alpha_5 L^5+\alpha_6 L^6)\ e(t)$ (5.26c).

Equations (5.26) and Table 5.5 deserve further comment. Firstly the exclusion of α_6 in Food seems odd but is explained by remembering that a reduction in the order of the M.A. involves the removal of two parameters: in this case α_6 and $e(-6)$. Because of the low t-value associated with $\hat{e}(-6)$ (0.26) it is not surprising that an overall test reveals that the order should be reduced. Secondly it was found that on statistical criteria (5.26c) should be retained for Miscellaneous. This conclusion is troubling as it would be better if there were no unit roots in the M.A. so that some other alternatives to (5.26c) were explored. These centered around the choice of differencing filter, and after initial failures with $(1-L)^2$ etc. it was decided to allow a non-unit root in the second part of the transfer function

i.e. $(1-L)(1-\beta L^4)\ y(t) = (1+\alpha_1 L+\alpha_4 L^4+\alpha_5 L^5+\alpha_6 L^6)\ e(t)$ (5.27).

Equations (5.26a), (5.26b) and (5.27) were then estimated and the resulting parameter estimates are set out in Table 5.6.

Table 5.6

Estimated Parameters of Equations (5.26a),(5.26b) and (5.27)(PH)

Series	$\hat{\alpha}_1$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{S}$
Food		-0.4363 (3.59)				87.60	6132
Appliances			-0.3135 (2.73)	-0.3933 (3.40)		44.66	2992
Miscella-neous	-0.1688 (1.21)	-0.8975 (7.08)	0.1429 (1.01)	0.1058 (1.01)	0.9229 (69.59)	352.92	22595

Generally the parsimonious models of (5.26a), (5.26b) and (5.27) are better than their counterparts in (5.25). Even now however the problem of near-unit roots in the M.A. remains for the Miscellaneous series but it was decided to fit these equations by the systems estimator; it may not matter that the M.A. roots are high in the light of the expository nature of the exercise and the high R^2 achieved.[9]

A final check on the adequacy of the maintained models is available from the correlograms of the residuals in Table 5.7.

---

9 Equation (5.27) exhibited convergence difficulties and it might have been better to estimate it as ARMA (1,4) rather than ARMA(1,6).

Table 5.7

Correlogram of the Residuals (8 lags)

Equations (5.26a), (5.26b) and (5.27)

Series	Lag								
	1	2	3	4	5	6	7	8	Q
1	-0.15	0.13	0.04	0.01	-0.07	-0.09	0.10	-0.06	5.08
	(1.31)	(1.19)	(0.42)	(0.10)	(0.64)	(0.78)	(0.83)	(0.56)	
3	0.16	-0.06	-0.16	-0.19	-0.08	-0.04	-0.07	0.04	7.91
	(1.40)	(0.51)	(1.39)	(1.65)	(0.68)	(0.35)	(0.61)	(0.39)	
4	-0.02	0.00	-0.03	-0.12	-0.17	0.07	-0.02	-0.16	5.66
	(0.16)	(0.01)	(0.25)	(1.05)	(1.50)	(0.61)	(0.19)	(1.39)	

There is strong evidence in Table 5.7 that the residuals are close to white noise and consequently (5.26a), (5.26b) and (5.27) may be regarded as an adequate specification for Food, Household Appliances and Miscellaneous respectively. In this form however there are a large number of parameters (24)- owing to the need to estimate e* - and to avoid computational difficulties with SYSTEM it was decided to impose the restriction e* = 0. The re-estimated equations are in Table 5.8.

Table 5.8

Estimated Parameters of Equations (5.26a), (5.26b) and (5.27) (PH/O)

Series	$\hat{\alpha}_1$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{S}$
Food		-0.3687				94.19	6970
Appliances			-0.3138	-0.3222		49.94	3649
Miscellaeous	-0.1763	-0.8929	0.1533	0.1106	0.9280	351.47	24603

A comparison of Tables 5.7 and 5.8 reveals:-

(i) The parameter magnitudes do not vary greatly if estimation is performed with PH/O rather than PH.

(ii) Letting $\hat{S}_1$ be the sum of squares associated with PH/O and $\hat{S}_2$ that associated with PH, under the null hypothesis that $e^*(j)$ $(j=1,\ldots,K)$ is zero the statistic $K^{-1}(\hat{S}_1-\hat{S}_2)/(N-M)^{-1}\hat{S}_2$ (where M is the total number of parameters fitted by PH) is distributed as $F_{K,N-M}$. The computed statistics were 2.39, 2.45 and 0.95 for Food, Appliances and Miscellaneous respectively so that only for Food is it likely that e^* is significantly different from zero.

Finally the set of equations was estimated by SYSTEM, although it was fairly obvious that the example was not a particularly good one in so far as there was little to be explained after differencing and the correlation between the residuals of different equations was weak. To appreciate the latter point, the covariance and correlation matrices were

$$\begin{bmatrix} 92.93 & 16.78 & 24.92 \\ 16.78 & 48.65 & 50.95 \\ 24.92 & 50.95 & 328.04 \end{bmatrix} \qquad \begin{bmatrix} 1.00 & 0.25 & 0.14 \\ 0.25 & 1.00 & 0.40 \\ 0.14 & 0.40 & 1.00 \end{bmatrix}.$$

5.5.2 Estimation

Both SYSTEM and the systems version of A.F.D. (called A.F.D./S) were applied to the selected equations and efficiency gains were tabulated for each estimator. Each table will contain:-

(i) The parameter estimates for SINGLE(A.F.D.) and SYSTEM(A.F.D./S).

(ii) The ratio of SINGLE(A.F.D.) to SYSTEM(A.F.D./S) variances (ϕ_1).

(iii) The ratio of asymptotic SINGLE(A.F.D.) to asymptotic SYSTEM(A.F.D./S) variances (ϕ_2). These were computed by inserting the relevant parameter estimates into (5.23).

Table 5.9

Efficiency Comparison of SYSTEM and SINGLE

Equations (5.26a), (5.26b) and (5.27)

Parameter	SINGLE	SYSTEM	ϕ_1	ϕ_2
$\hat{\alpha}_{14}$	-0.3687	-0.4342	1.22	0.72
$\hat{\alpha}_{25}$	-0.3138	-0.3006	1.25	1.06
$\hat{\alpha}_{26}$	-0.3222	-0.3475	1.24	1.09
$\hat{\beta}_{31}$	0.9280	0.9284	1.14	1.10
$\hat{\alpha}_{31}$	-0.1763	-0.1840	1.21	0.94
$\hat{\alpha}_{34}$	-0.8929	-0.8928	1.18	1.00
$\hat{\alpha}_{35}$	0.1533	0.1635	1.20	1.06
$\hat{\alpha}_{36}$	0.1106	0.1101	1.17	1.03

There are a number of points emerging from Table 5.9 that demand attention. Firstly there are only minor changes in the parameter magnitudes in the transition from one estimator to the other. To test whether there has been a significant change denote SYSTEM estimates by $\hat{\theta}_1$ and SINGLE estimates by $\hat{\theta}_2$: then the statistic $(\hat{\theta}_1-\hat{\theta}_2)^T V^{-1}_{\hat{\theta}_1\hat{\theta}_1}(\hat{\theta}_1-\hat{\theta}_2)$ is distributed as $\chi^2(k)$ (k is the dimension of $\hat{\theta}_1$ and $\hat{\theta}_2$) under the null hypothesis, and for the comparison the calculated statistic was 7.2 which would only be significant at a level greater than 0.5. Secondly the efficiency gains are, as was to be expected, only moderate. Two measures of this were adopted:

(i) $\text{tr}\,(V_{\hat{\theta}_2\hat{\theta}_2})/\text{tr}(V_{\hat{\theta}_1\hat{\theta}_1})$

(ii) $\det(V_{\hat{\theta}_2\hat{\theta}_2})/\det(V_{\hat{\theta}_1\hat{\theta}_1})$.

The first of these was 1.22 and the second 4.99. Thirdly the empirical and asymptotic variances are close for all parameters but β_{31}, and in this case the discrepency is so great as to cast some doubt upon the estimate of the covariance matrix obtained from (5.11) in a situation of small samples and high roots.

A comparison between A.F.D. and A.F.D./S was performed on the above lines and Table 5.10 corresponds to Table 5.9.

Table 5.10

Efficiency Comparison of A.F.D. and A.F.D./S

Equations (5.26a),(5.26b) and (5.27)

Parameter	A.F.D.	A.F.D./S	ϕ_1	ϕ_2
$\hat{\alpha}_{14}$	-0.4702	-0.5259	1.26	0.86
$\hat{\alpha}_{25}$	-0.3025	-0.1860	1.34	1.11
$\hat{\alpha}_{26}$	-0.3211	-0.4602	1.32	1.08
$\hat{\beta}_{31}$	0.9187	0.9417	1.34	1.14
$\hat{\alpha}_{31}$	-0.4063	-0.4097	1.20	1.17
$\hat{\alpha}_{34}$	-0.2978	-0.3156	1.33	1.12
$\hat{\alpha}_{35}$	-0.0041	-0.0377	1.23	1.18
$\hat{\alpha}_{36}$	0.1740	0.1902	1.23	1.10

The conclusions concerning A.F.D./S vis à vis A.F.D. are similar to those for SINGLE and SYSTEM except that the efficiency gains would seem to be larger in the frequency domain - the most prominent indicator being the very high determinantal ratio. Otherwise the main item of interest in Table 5.10 is the wide discrepency in estimates of the M.A. parameters for Miscellaneous between A.F.D. and SYSTEM (PH): no doubt caused by the unit roots. Computationally A.F.D./S is inferior to SYSTEM with 1 minute being required for convergence of SYSTEM (from arbitrary starting values) and 5 minutes for A.F.D./S (from the single equation estimates). However, although high, it does seem as if the computational cost would not be excessive for small systems.[10]

10 A Newton-Raphson version was written for both estimators and for the system of the text it required 5 minutes (SYSTEM) and (probably) 40 minutes (A.F.D./S) for convergence : the latter figure being only an estimate based on the computation time for one iteration. As remarked earlier the inordinate amount of computer time was a concomitant of the function evaluations required for second derivatives.

5.6 Conclusion

Chapter 5 has shown that it is both possible and beneficial to develop estimators for systems of equations in which there are M.A. disturbance terms. Although limited, the Monte Carlo and empirical experiments were sufficiently encouraging in their conclusions to justify the application of systems estimators when the situation so demands. As there are now a number of fields of econometric research e.g. inter-related factor demands and systems of demand equations, in which a non-simultaneous set of equations arise, and in which there has recently been a substantial use of quarterly data [76], the likelihood that the estimators discussed in this chapter will be required is becoming greater.

Appendix 9

Derivation of the Asymptotic Covariance Matrix for the M.L. Estimator of Systems of Equations

By an application of the methodology of Appendix 3 the asymptotic covariance matrix of $\sqrt{N}\ (\hat{\theta}-\theta)(\theta=\begin{bmatrix}\beta\\ \alpha\end{bmatrix})$ is

$$V_{\hat{\theta}\hat{\theta}} = \underset{N\to\infty}{\text{Lim}}\ [E(\frac{2\pi}{N}\ \frac{\partial\varepsilon^{T}(\lambda)}{\partial\theta}\ (\Omega^{-1}\otimes I)\ \frac{\partial\bar{\varepsilon}(\lambda)}{\partial\theta})]^{-1} \qquad (A9.1).$$

Consider the cross derivative between the d'th A.R. parameter in the r'th equation and the g'th A.R. parameter in the s'th equation i.e. β_{rd} and β_{sg}. From (A9.1) we seek

$$\underset{N\to\infty}{\text{Lim}}\ \frac{1}{N}\ E\ [2\pi\ \frac{\partial\varepsilon^{T}(\lambda)}{\partial\beta_{rd}}\ (\Omega^{-1}\otimes I)\ \frac{\partial\bar{\varepsilon}_{s}(\lambda)}{\partial\beta_{sg}}\ ] \qquad (A9.2).$$

If neither β_{rd} nor β_{sg} are common to other equations the cross term may be replaced by

$$\underset{N\to\infty}{\text{Lim}}\ \frac{1}{N}\ E(2\pi\ \sum_{j=1}^{N}\ \frac{\partial\varepsilon_{r}(\lambda_{j})}{\partial\beta_{rd}}\ \sigma^{rs}\ \frac{\partial\bar{\varepsilon}_{s}(\lambda_{j})}{\partial\beta_{sg}}\ ) \qquad (A9.3).$$

Now $\varepsilon_{r}(\lambda_{j}) = \dfrac{Y_{r}(\lambda_{j})\ B_{r}(\lambda_{j})}{A_{r}(\lambda_{j})}$ so that

$$\frac{\partial\varepsilon_{r}(\lambda_{j})}{\partial\beta_{rd}} = \frac{Y_{r}(\lambda_{j})\ e^{id\lambda_{j}}}{A_{r}(\lambda_{j})} \qquad (A9.4)$$

$$\frac{\partial\bar{\varepsilon}_{s}(\lambda_{j})}{\partial\beta_{sg}} = \frac{\bar{Y}_{s}(\lambda_{j})\ e^{ig\lambda_{j}}}{\bar{A}_{s}(\lambda_{j})} \qquad (A9.5).$$

Substituting (A9.5) and (A9.4) into (A9.3) we obtain

$$\lim_{N\to\infty} \frac{1}{N} E\left(2\pi\, \sigma^{rs} \sum_{j=1}^{N} \frac{I_{rs}(\lambda_j)\, e^{i(d-g)\lambda_j}}{A_r(\lambda_j)\, \bar{A}_s(\lambda_j)}\right) \qquad \text{(A9.6)}.$$

Asymptotically the expectation of the cross periodogram is the cross spectrum ([40 p.213]) so that

$$\lim_{N\to\infty} E(I_{rs}(\lambda_j)) = \frac{\sigma_{rs}\, A_r(\lambda_j)\, \bar{A}_s(\lambda_j)}{2\pi\, B_r(\lambda_j)\, \bar{B}_s(\lambda_j)} \quad ,$$

and (A9.6) simplifies to

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i(d-g)\lambda}\, \sigma_{rs}\, \sigma^{rs}}{B_r(\lambda)\, B_s(\lambda)}\, d\lambda \qquad \text{(A9.7)}$$

which agrees with (5.23) of the text. By a similar treatment of the other cross terms it may be shown that the remaining expressions in (5.23) hold as well.

CHAPTER 6

Parametric Estimation of Some Time Series Models

6.1 Introduction

Earlier chapters have centred on the derivation and estimation of ARMAX equations with particular attention being given to the sampling distributions of the proposed estimators. In the final two chapters some applications of the techniques given in Chapter 2 will be outlined and an assessment will be made of their usefulness to econometricians. The present chapter will be concerned with a class of models, involving the decomposition of a time series into mutually orthogonal elements, that has appeared in various guises in the econometric literature, and it will be seen that with some modification these models may be estimated by the algorithms of Chapter 2.

A brief review of the chapter follows. Section 6.2 outlines the model to be employed and reference will be made to four areas of applied research in which it arises.[1] The following section selects some of the simplest forms of the model for study, various estimators are presented, and a distinction based on increasing efficiency is made.[2] This leads into some Monte Carlo studies illustrating (and evaluating) the efficiency of the various algorithms of Section 6.4. Finally 6.5 applies the methods to a number of economic time series in order to indicate areas in which the approach has potential.

6.2 A Time Series Model

A major proposition of this chapter is that a number of (seemingly unrelated) problems arising in the analysis of time series

---

[1] A fifth is recounted in Chapter 7.

[2] Efficiency in the sense of minimum variance.

may be regarded as having come from a set of equations

$$y(t) = T(t) + I(t) \tag{6.1a}$$

$$T(t)\ B(L) = e_1(t) \tag{6.1b}$$

$$I(t) = e_2(t) \tag{6.1c},$$

where $B(L) = 1-\beta_1 L- \ldots -\beta_p L^p$ satisfies the usual root requirements and[3]

$$E(e_k(t)) = 0 \quad k = 1,2$$

$$E(e_k(t)\ e_\ell(t-j)) = \sigma_{kk} \qquad k = \ell = 1,2; \quad j = 0$$

$$= 0 \qquad k,\ell = 1,2; \quad j \neq 0.$$

The reduced form of (6.1a) - (6.1c) is

$$B(L)\ y(t) = e_1(t) + B(L)\ e_2(t) \tag{6.2}.$$

With the aid of equations (6.1a) - (6.2) it will be possible to demonstrate that four problems arising in econometrics and statistics have a common structure. The four are:-

(i) A.R. with superimposed noise.

(ii) Permanent and Transitory Components of a series.

(iii) Seasonal Components of a series.

(iv) Adaptive Forecasting.

---

[3] The assumption of uncorrelated disturbances will be retained throughout the chapter. Therefore whenever a disturbance is subscripted it should be understood that it is uncorrelated with all other subscripted disturbances. One might also interpret the model in (6.1a)-(6.1c) as a more general one of "signal embedded in noise" where a specific form viz an A.R. is assumed for the signal T(t).

(i) A.R. with Superimposed Noise

Let us assume that a variable T(t) follows a p'th order A.R. but that it is observed with an error. Under the assumption that the observation error is i.i.d. and uncorrelated with the error arising from the A.R., a model compatible with (6.1a)-(6.1c) is generated. In this guise statisticians have researched the estimation of the parameters of B(L) for a number of years, with Walker [111], Parzen [89] and Bailey [5] being notable contributors. Their interest has been stimulated by the belief that a famous historical series - Wolfer's sunspot observations - follows a second order A.R. with errors of observation e.g. Yule [122], but as well Walker acknowledges that ".... further investigations ..... should be of practical importance". [111 p.33].

(ii) Permanent and Transitory Components of a Series

Friedman [28], in the course of explaining the constancy of the long run marginal propensity to consume in the United States, proposed that income and consumption could be viewed as the summation of two orthogonal components: one of which (termed transitory) was to be white noise, whereas the other (permanent) was to be autocorrelated. Obviously the system of equations (6.1a)-(6.1c) will represent such a scheme if, when y(t) is income, T(t) is designated permanent income.[4]

What form should B(L) take? A tempting answer is

$$B(L) = 1 - \beta L \qquad (6.3).$$

---

[4] Some discussion has arisen around the nature of the transitory component under the heading of windfall gains and it appears that the specification of the transitory component as white noise may be inaccurate.

With B(L) as in (6.3) (6.2) becomes

$$y(t)(1-\beta L) = e_1(t) + e_2(t)(1-\beta L) \qquad (6.4),$$

or by noting that the right hand side of (6.4) has the correlogram of a first order M.A.

$$y(t)(1-\beta L) = (1+\alpha L)\,\varepsilon(t) \qquad (6.5).$$

As Friedman realised the conceptual distinction employed in decomposing income is of little value unless it may be translated into empirical constructs i.e. an estimate of T(t) must be formed. To construct such an estimate we may draw upon the well developed body of theory concerned with the optimal extraction of a signal (T(t)) from a series (y(t)) that is contaminated with noise (I(t)). The fundamental theorem to carry out the extraction (Whittle [117])is:

<u>If a series y(t)(t=1,...,N) is a linear combination of a signal T(t) and noise I(t) and can be expressed as $y(t) = \Phi(L)\varepsilon(t)$ where $\varepsilon(t)$ is i.i.d. $(0,\sigma^2)$ then the best linear extraction of T(t) is given by</u>

$$\hat{T}(t) = \frac{1}{\sigma^2\,\Phi(L)} \left[\frac{\Gamma_{Ty}(L)}{\Phi(L^{-1})} \right]_+ y(t)$$

where + indicates that only positive powers of L are retained. For the special case when $B(L) = 1-\beta L$, $\Phi(L) = \frac{1+\alpha L}{1-\beta L}$ Nerlove [81] has shown that the optimal estimate is

$$\hat{T}(t) = \frac{\beta+\alpha}{\beta} \quad \frac{y(t)}{1+\alpha L} \qquad (6.6).$$

Rearranging (6.6) and setting $A = \alpha(\frac{1}{\beta} - 1)$, $B = (\frac{\beta+\alpha}{\beta})$ yields

$$\hat{T}(t) = (1-B+A)\,\hat{T}(t-1) + By(t) \qquad (6.7).$$

(6.7) shows that the optimal estimate of permanent income would be derived from a recursive calculation and by comparing it to Wright [120 eq. 3 p.845] it is plain that (6.7) is the form used by Friedman in constructing his estimate of permanent income. Therefore under the assumption that permanent income follows a first order A.R. it has been shown that an optimal estimate may be formed by an exponential smoothing of all past values of the current income series (see (6.6)).

It is of some interest to dwell on (6.7) and the last statement. Firstly Holmes [51] has recently claimed that an exponential weighting contradicts the assumption of uncorrelated components; yet this condition was an essential step in the derivation of (6.6).[5] The paradox is resolved by noting that Holmes derives his "impossibility theorem" from the assumption that y(t) is an i.i.d. variable, and an examination of the theoretical covariance function of y(t) shows that this is invalid.[6] Secondly a comparison of Friedman's parameters (A=0.02, B=0.4) with those of (6.6) yield β=1.3, α=-0.62, so that, although the weighting function derived from the simple model is identical to Friedman's, the magnitude of β is at variance with the restrictions of Chapter 1. Therefore if permanent income is to be generated by a system such as (6.1a)-(6.1c) it will be mandatory to

---

5 Only if E(T(t)I(t))= 0 is it permissible to write $\Gamma_{Ty} = \Gamma_{TT}$ - a basic relationship required in the proof.

6 An intuitive explanation comes from recognizing that the very notion of a permanent component in a series carries with it the necessity of some autocorrelation in that series. If the series really were i.i.d. then the most reasonable assumption to be made would be that there is only a transitory component.

determine the extent to which the restricted parameter space will generate "realistic"series. This point is touched on later in the chapter. Lastly, ignoring for the moment questions arising from the root restrictions, the fact that the generated model is ARMA insinuates that estimates of permanent components might be constructed from the parameter estimates $\hat{\beta}$, $\hat{\alpha}$ obtainable from fitting ARMA models to the data.

(iii) Seasonal Components of a Series

Some authors have posited that time series might be regarded as the summation of various orthogonal elements viz. trend-cycle, seasonal and irregular. Indeed, as Grether has remarked [35], this subjective decomposition has a long and honourable tradition in economic methodology and still forms the basis of present-day methods of seasonal adjustment. In practice most controversy has arisen from the fact that it is usually not very clear or explicitly indicated just what is assumed about the nature of these components, with the result that any description of the components is inevitably a function of the observer, and the best one might hope for would be a clear definition in each case.

Nerlove, Couts and Grether [18] have been the major proponents of (6.1) as a description of time series. Their analysis would identify T(t) as the trend and I(t) as the irregular. With the addition of a third factor (the seasonal S(t)) they found that it was possible to obtain good forecasts for the series Unemployed Males, aged 14-19 from January 1948 to December 1961; the model employed for this purpose being

$$y(t) = T(t) + I(t) + S(t)$$

$$T(t)\ (1-\beta L)^2 = e_1(t)$$

$$S(t)\ (1-L^{12})^2 = e_2(t) \qquad (6.8).$$

$$I(t) = e_3(t)$$

Again it is clear that an ARMA model is the reduced form and defining $z(t) = (1-L^{12})^2\ y(t)$ it will be ARMA (2,26) in z(t).

(iv) Adaptive Forecasting

As well as designing an optimal filter to extract T(t), economists may wish to forecast y(t), and provided the series y(t) has a moving average representation the weighting function to be applied to y(t) to generate optimal forecasts is equation (1.23) (p.25). Specializing this formula to the simple model with reduced form (6.5), one-period ahead forecasts of y(t+1) at time t (designated y*(t)) may be made from

$$y^*(t) = \frac{\beta+\alpha}{1+\alpha L}\ y(t) \qquad (6.9),$$

or

$$y^*(t)\ (1+\alpha L) = (\beta+\alpha)\ y(t) \qquad (6.10).$$

Now consider the case $\beta=1$.[7] (6.10) then reduces to

$$y^*(t) = -\alpha y^*(t-1) + (1-\alpha)\ y(t) \qquad (6.11)$$

which (with $A=-\alpha$) is easily recognizable as the exponentially weighted or adaptive forecast of Holt [52] (see [117 p.97] for a proof) in that

---

[7] Hannan [42] presents a rigorous proof that Whittle's optimal prediction and extraction formulae are applicable in this case even though the series is no longer stationary.

Holt's formula provides a one-period ahead forecast of (say) sales by a linear combination of the actual sales in the current period and the forecast of sales made for the current period in the previous period. Wagle, in a recent survey of short-term forecasting techniques, has asserted that (6.11) is "appropriate for products with stable sales rates and little seasonal influences"[109 p243]thereby justifying the trend plus irregular interpretation given here. A similar conclusion exists for higher order differences in the trend e.g. Theil and Wage [104] have derived the adaptive forecasting formula for a series with trend and irregular factors when the trend follows $T(t)(1-L)^2 = e_1(t)$. As with the previous three sub-sections it is obvious that the parameters of adaptive forecasting formulae may be identified with the coefficients of an ARMA model.

6.3 Parametric Estimation of Some Simple Models

What algorithms are available for the estimation of the parameters of (6.1)? An answer from management science is that ".... selection of the unknown parameters .... are usually obtained by trial and error" [109] but a more optimistic one is to be found in the literature concerned with the design of optimal filters for signal extraction. Most of this appears in engineering journals, and although general solutions have been proposed (Yaglom [121]),the analytical difficulties have been so great as to restrict their use to very simple models.[8]

---

8 Scadding [97] reviews the significance of this work for economists and outlines the optimal filter when $B(L) = 1-\beta L$. A quick perusal of his paper reveals that extending the analytic approach would require great mathematical skill so that a numeric alternative would be valuable.

To date four simple methods have been proposed - those of Walker [110], Bailey [5],Leenders [65] and Grether [35]. As all are founded on the correlogram of the series it is instructive to consider some special cases. Letting $u(t) = B(L)e_2(t) + e_1(t)$ in (6.2), Table 6.1 contains the non-zero covariances of u(t) for a number of different B(L).

Table 6.1

Non-zero Covariances of u(t) for some special forms of B(L)*

Model	B(L)	γ_{uu}
I	$1-\beta L$	$\gamma_{uu}(0) = \sigma_1^2(1+(1+\beta^2)\lambda)$
		$\gamma_{uu}(1) = -\sigma_1^2\lambda\beta$
II	$1-L$	$\gamma_{uu}(0) = \sigma_1^2(1+2\lambda)$
		$\gamma_{uu}(1) = -\sigma_1^2\lambda$
III	$(1-L)^2$	$\gamma_{uu}(0) = \sigma_1^2(1+6\lambda)$
		$\gamma_{uu}(1) = -4\sigma_1^2\lambda$
		$\gamma_{uu}(2) = \sigma_1^2\lambda$

* λ is the variance ratio σ_2^2/σ_1^2.

With the aid of the covariance functions in Table 6.1 the four methods mentioned above will be explained along with two new algorithms.

(i) Walker/Leenders Estimators

Both estimators rely on an equation of theoretical and empirical

covariances of u(t).[9] Take Model II as an example. It is possible to form estimates of $\hat{\gamma}_{uu}(0)$ and $\hat{\gamma}_{uu}(1)$ from the differenced data, and then obtain estimates of σ_1^2 and λ by equating the theoretical values of $\gamma_{uu}(0)$ and $\gamma_{uu}(1)$ (given in Table 6.1) to the empirical values. A similar procedure applies to Model III. For Model I there is an added complexity in the presence of β in both the A.R. and M.A., Walker argues that instrumental variables be applied to the first order A.R. to yield a consistent estimate of β, and then the consistent estimates of $\gamma_{uu}(0)$ and $\gamma_{uu}(1)$ are utilized to find $\hat{\sigma}_1^2$ and $\hat{\lambda}$.

Walker provides a general proof that the estimator is consistent and that $N^{\frac{1}{2}}(\hat{\lambda} - \lambda \; ; \; \sigma_1^2 - \sigma_1^2)$ will have, as $N \to \infty$, a multinormal distribution with zero mean. Furthermore he derives the efficiency of the estimator vis à vis the M.L. estimator and finds that efficiency declines dramatically for (6.5) as β and $|\alpha|$ rise. This result is sensible in that as $|\alpha|$ becomes smaller the sources of inefficiency will be progressively removed. From this it emerges that Walker's method is not robust against many of the models that may arise, and Leenders attempted to improve upon it by making corrections to the estimates obtained.[10] Although he seems to have been successful for the special case of Model III the analytical difficulties inherent in his derivation

---

9 Leenders' paper was not available to the author and the source of information regarding it was Wagle. The skimpy account there precludes any detailed discussion of it in this thesis.

10 It is not really Walker's method as he was only assessing the efficiency of it without advocating its use.

must reduce its utility for higher order models (particularly if there are unknown roots in B(L)).

There are further deficiencies in the "correlogram estimator" of Walker's article. One of the more important is that in small samples there is no guarantee of the root restrictions being satisfied by the instrumental variable estimator of B(L). In some Monte Carlo studies constructed to assess the efficiency of this estimator there were numerous instances of the roots of $\hat{B}(L) = 0$ lying inside the unit circle. Similarly it was possible to obtain estimates of λ that were negative, and this behaviour was found in a number of empirical examples.

Finally Walker's paper is of wider interest than the examination of a particular estimator in as much as it contains the asymptotic covariance matrix of the efficient estimator of β and λ in Model I i.e.

$$N \text{ var } (\hat{\beta}) = \frac{(1-\beta^2)(1-\alpha)^2}{(\beta-\alpha)^2} \qquad (6.12)$$

$$\alpha^4 N \text{ var}(\hat{\lambda}) = \lambda(1-\alpha)\{(3\beta-4\beta^3)\alpha^5 + \alpha^4+2\beta\alpha^3 + (2-4\beta^2)\alpha^2-\alpha\beta+1\} \qquad (6.13)$$

where α is the M.A. parameter in (6.5) and is a root of the equation

$$\lambda\beta(L^2+1) - L(1+\lambda(1+\beta^2)) = 0 \qquad (6.14).$$

(6.12) and (6.13) may be used to evaluate the sampling performance of any proposed estimator and will be referred to in some Monte Carlo studies in the following section.

(ii) Grether

Couts, Grether and Nerlove [18] (and later Grether in his doctoral thesis [35]) have advocated"...the provisional procedure of generating theoretical spectra for the process ... using various

values of the parameters. By comparing these with the spectrum estimated from the corresponding series derived from the series to be predicted we were able to select values... that provided a close match between the empirical and theoretical spectra" [18 p.19]. There is little that can be said about the statistical properties of such an ad hoc approach, but as it is known that the estimated spectrum may diverge quite substantially from its true shape the degree of confidence to be placed in any estimates must be low. On the other hand some check is available in that the forecast and actual series may be compared for a number of parameter sets and the coefficients minimizing the sum of squared deviations selected: in this way the technique resembles a grid search. An attempt was made to formalize the method by devising an algorithm that would minimize the squared deviations between the theoretical and empirical spectra by generating corrections to parameter guesses, but this was not very successful.

Nevertheless matching of spectra has some applicability. Taking Model II as an example, Table 6.1 shows that the first autocorrelation coefficient of u(t) must be negative implying that $\alpha_1 < 0$ and thereby circumscribing the spectral shape that the differenced series may possess. Therefore a visual matching of the estimated and theoretical spectra provides important evidence of the potential adequacy of any Model e.g. if after differencing $f_{uu}(\lambda)$ had greater power at the zero frequency than at π it is doubtful if Model II would be appropriate, because a first order M.A. with $\alpha_1 < 0$ has greatest power at π.

(iii) An ARMA Model Estimator

A solution hinted at already is to apply ARMA estimators to Models I-III. Probably this will be more efficient than a correlogram

estimator as it has been mentioned (Section 1.5.3) that estimates of the M.A. parameters from solving the covariance equations are very inefficient, whereas the estimate of α obtained by the M.L. methods of Chapter 2 was not. To ensure fully efficient estimates all restrictions must be incorporated into the estimation process. For the models of Table 6.1 they are analytically derivable from the covariance functions as

(a) Model I $\quad \gamma_{uu}(1) > 0 \quad$ if $\beta < 0$

$\gamma_{uu}(1) < 0 \quad$ if $\beta > 0$

(b) Model II $\quad \gamma_{uu}(1) < 0$

(c) Model III $\quad \gamma_{uu}(1) = -4\gamma_{uu}(2)$

$\gamma_{uu}(1) < 0$

$\gamma_{uu}(2) > 0$.

Translated into restrictions on the parameters of the corresponding ARMA equations (and adopting the general ARMA nomenclature) these are

(a) $\alpha_1 > 0 \quad$ if $\beta < 0$

$\alpha_1 < 0 \quad$ if $\beta > 0$

(b) $\alpha_1 < 0$

(c) $\alpha_1(1+\alpha_2) = -4\alpha_2$

$\alpha_1 < 0$

$\alpha_2 > 0$.

Despite the fact that a restricted ARMA estimator is fully efficient the determination of restrictions becomes increasingly cumbersome as the order of B(L) grows, and there is the allied difficulty that the sign and equality restrictions call for a programming approach. A possible solution would be to ignore the restrictions. Although this may be reasonable for Models I-III, as the order of B(L) grows the number of parameters to be estimated rises at double the rate, resulting in a loss of valuable degrees of freedom. Nevertheless, as a first stage estimator it has much to recommend it and was found to be useful in a number of applications. In later discussion the estimator is designated as PH/U i.e. the unrestricted Phillips estimator.

(iv) Covariance Factorization and an ARMA Estimator

This algorithm builds on (iii) but incorporates all constraints numerically. To illustrate the nature of the solution the general model in (6.1a)-(6.1c) is adopted and the reduced form, (6.2), will be re-written as

$$B(L)\, y(t) = u(t) = B(L)e_2(t) + e_1(t) \qquad (6.15).$$

Four self evident propositions are required in the subsequent analysis.

A. The series u(t) has a M.A. form ($=A(L)\varepsilon(t)$) of order p and this establishes the equivalence

$$\Gamma_{uu}(L) = \sigma_1^2[\lambda B(L)B(L^{-1})+1] \equiv \sigma_\varepsilon^2[A(L)A(L^{-1})].$$

B. The roots of the covariance function $\Gamma_{uu}(L)$ obtained from both sides of the equivalence relation are equal.

C. As $\Gamma_{uu}(L)$ is a covariance function there will be p roots inside and p roots outside the unit circle.

D. The set of all roots outside the unit circle may be identified as the roots of $A(L) = 0$ (Condition 7 of Chapter 1).

Propositions A-D indicate that the parameters of $A(L)$ may be found by factoring the $\Gamma_{uu}(L)$ polynomial into its roots and extracting those that lie outside the unit circle.[11] This then becomes the basis for the following algorithm:-

(a) Assume initial estimates $\lambda^{(0)}$, $\beta^{(0)}$ of the unknown parameters. Later we will discuss the origin of these.

(b) Construct an estimate of the covariance function $\Gamma_{uu}(L)$ from $\lambda^{(0)}$, $\beta^{(0)}$

$$\text{i.e. } \Gamma_{uu}{}^{(0)}(L) = \sigma_1^2[1+\lambda^{(0)}B^{(0)}(L)B^{(0)}(L^{-1})].$$

(c) Solve for the roots of $\Gamma_{uu}{}^{(0)}(L)$. Selecting those that lie outside the unit circle $(\phi_1{}^{(0)},\ldots,\phi_p{}^{(0)})$ we may construct an estimate of $A^{(0)}(L)$ from

$$A^{(0)}(L) = \prod_{j=1}^{p}(1-\phi_j{}^{(0)-1}L).$$

(d) Using the relation $u^{(0)}(t) = A^{(0)}(L)\varepsilon^{(0)}(t)$, $\varepsilon^{(0)}(t)$ is generated recursively (assuming $\varepsilon(-1)=0$).

(e) Once $\varepsilon^{(0)}(t)$ is determined it is possible to construct estimates of the derivatives $\frac{\partial\varepsilon(t)}{\partial\lambda}$, $\frac{\partial\varepsilon(t)}{\partial\beta}$ (at $\lambda^{(0)}$, $\beta^{(0)}$) by perturbating β and λ and then using difference formulae. A vector of corrections to $(\lambda^{(0)}, \beta^{(0)})$ may be found from the Gauss-Newton algorithm i.e.

---

11 The article by Nerlove [81] is a good reference to this procedure (particularly the Appendix).

letting $\theta^{T(n)} = (\lambda^{(n)} ; \beta^{(n)})$ and

$\frac{\partial \varepsilon}{\partial \theta} = \left[\frac{\partial \varepsilon}{\partial \lambda} ; \frac{\partial \varepsilon}{\partial \beta} \right]$ (where all partial derivations are evaluated at $\theta^{(n)}$) this is

$$\theta^{(n)} - \theta^{(n-1)} = - \left(\frac{\partial \varepsilon^T}{\partial \theta} \frac{\partial \varepsilon}{\partial \theta} \right)^{-1} \frac{\partial \varepsilon^T}{\partial \theta} \varepsilon \qquad (6.16).$$

Clearly there are similarities between this algorithm (called PH/R i.e. the restricted Phillips estimator) and PH/O in that there is an identical set of estimating equations once A(L) is determined. It should be noted that all restrictions from the covariance functions have been utilized in the construction of $\varepsilon(t)$ so that unlike the ARMA estimator of the preceding sub-section it should be fully efficient. Furthermore it has the advantage that it is easily extended to three-components models - although no formal (theoretical) demonstration is presented. Finally in practice e* would be treated as a parameter to be estimated along with the others and initial values of λ and β would be selected along the lines of 2.6.1(p. 70).

The following example illustrates the performance of the algorithm on simulated data. For Model I with $\beta = 0.8$, $\lambda = 1.3$ and a sample size of 75, the iterations proceed as in Table 6.2. Arbitrary starting values were selected and these are iteration 0.

Table 6.2

An Example of the Covergence of PH/R

Iteration No.	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\alpha}_1$	$S = \Sigma\varepsilon^2(t)$
0	4.00	0.50	-0.096	301.3
1	1.91	0.56	-0.180	296.7
2	1.02	0.73	-0.316	274.5
3	1.65	0.81	-0.263	265.1
4	1.65	0.81	-0.263	265.1

Convergence was achieved in 9 secs. on the I.B.M. 360/50. During the construction of the algorithm a number of computational difficulties arose and a full description of these (and the solutions) is in Appendix 10. Most of the problems revolved around the failure of standard programs such as Bairstow's to find the roots of $\Gamma_{uu}(L)$ and eventually, echoing a discovery by [36], Muller's method was selected. As will become evident in the following sections however even this was not entirely satisfactory.

(v) Bailey's Estimator

Bailey is interested in optimal prediction rather than estimation but his mode of attack has some similarities to (ii) and (iv). Appendix 11 shows that the weights to be applied to y(t) to generate optimal forecasts $(\omega_o,\ldots,\omega_\kappa)$ may be determined from

$$\omega = W^{-1}w \qquad (6.17),$$

where $|H(\lambda)|^2 = (2\pi)^{-1}\,(\sigma_2^2|B(\lambda)|^2 + \sigma_1^2)$

and $W(k,\ell) = |H(\lambda)|^2 \cos(k-\ell)\lambda \qquad k,\ell = 0,\ldots,\kappa$

$w(k) = |H(\lambda)|^2 \cos k\lambda \qquad k = 0,\ldots,\kappa.$

Now it is easily seen that $|H(\lambda)|^2$ is the spectrum of $u(t) = B(L)e_2(t) + e_1(t)$ (in (6.2)) so that it may be represented by a p'th order M.A. i.e. $u(t) = A(L)\varepsilon(t)$. By guessing the roots of A(L) (mainly by guidance from the likely roots of B(L)) it is possible to construct optimal predictions of y(t) and thereby minimize the forecast variance. There is no systematic updating of estimates however and the algorithm merely degenerates into a grid search for isolating the roots of A(L) yielding minimum mean square error forecasts. Although this is a major disadvantage of the procedure there is an offsetting gain in that it is not necessary to impose the restriction $E(T(t)I(t)) = 0$ (although this was done in (6.17)).

6.4 Monte Carlo Studies

Three models were selected for sampling experimentation viz. Models I, II and IV.

Model IV

$$\begin{aligned} y(t) &= T(t) + S(t) + I(t) \\ T(t)(1-L) &= e_1(t) \\ S(t)(1-0.9L^4) &= e_2(t) \\ I(t) &= e_3(t) \end{aligned} \qquad (6.18).$$

Model IV may be regarded as representative of a quarterly series with a trending mean, a stationary seasonal pattern, and random irregular movements. The reduced form of (6.18) is

$$(1-L)(1-0.9L^4)y(t) = (1+\alpha_1 L+\alpha_2 L^2+\alpha_3 L^3+\alpha_4 L^4+\alpha_5 L^5)\varepsilon(t) \qquad (6.19).$$

Table 6.3 includes the values of λ chosen (or λ_1 and λ_2 in the case of Model IV) and the implied M.A. parameters for Models I, II and IV.

Table 6.3

Variance Ratios and the Implied M.A. Parameters of Models I, II and IV

Model	λ_1	λ_2	B(L)	A(L)
I	10		$(1-0.9L)$	$(1-0.7062L)$
II	4		$1-L$	$(1-0.6096L)$
IV	1	1	$(1-L)(1-0.9L^4)$	$(1-0.5688L-0.0457L^2-0.0647L^3-0.4666L^4+0.1919L^5)$

The rationale for picking these Models was as follows. Firstly efficiency comparisons may be performed on Models I and II, as Walker has provided the asymptotic variances for the former and there is only a sign restriction preventing PH/U from being fully efficient in the case of the latter. For the particular M.A. parameter of Model II it is unlikely that this condition will be violated so that PH/U provides something in the nature of a control solution. Both Models I and II are suitable for yearly data (and will be so used later) whilst Model IV is explicitly adopted to simulate quarterly time series. Tables 6.4, 6.5 and 6.6 contain the summary statistics with only the M.A. and A.R. parameters presented (for reasons to be discussed later).

Table 6.4

Summary Statistics for PH/U and PH/R

Models I, II, and IV: Sample Size N = 40

PART A

Model	P	PH/U			PH/R		
		M	B	B/SE	M	B	B/SE
I	0.9	0.7814	-0.1186	2.71	0.7852	-0.1148	3.16
	-0.7062	-0.6361	0.0701	1.24	0.6083	0.0979	2.21
II	-0.6096	-0.6307	-0.0217	0.79	-0.6265	-0.0169	0.64
IV	0.9	0.7310	-0.1690	5.47	0.7685	-0.1315	8.09
	-0.5688	-0.5022	0.0666	2.16	-0.4951	0.0737	2.90
	-0.0457	-0.0029	0.0428	1.06	-0.0685	-0.0228	0.90
	-0.0647	-0.0002	0.0645	1.48	-0.0646	0.0001	0.02
	-0.4666	-0.4006	0.0660	1.31	-0.3456	0.1210	5.42
	0.1919	0.1138	-0.0781	2.16	0.0932	-0.0987	5.39

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
I	AR	0.3097	2.39	0.3316	0.2565	1.98	0.2810
	MA	0.3985		0.4046	0.3138		0.3287
II	MA	0.1953	1.56	0.1965	0.1876	1.50	0.1884
IV	AR	0.2185		0.2196	0.1149		0.1746
	MA	0.2181		0.2280	0.1794		0.1939
	MA	0.2844		0.2876	0.0667		0.0705
	MA	0.3077		0.3144	0.0358		0.0358
	MA	0.3571		0.3631	0.1580		0.1990
	MA	0.2552		0.2669	0.1294		0.1627

Table 6.5

Summary Statistics for PH/U and PH/R

Models I,II, and IV ; Sample Size N = 70

PART A

Model	P	PH/U			PH/R		
		M	B	B/SE	M	B	B/SE
I	0.9	0.8018	-0.0982	2.38	0.8267	-0.0733	2.83
	-0.7062	-0.6344	0.0718	1.38	-0.6544	0.0518	1.44
II	-0.6096	-0.6270	-0.0174	0.95	-0.6271	-0.0175	0.97
IV	0.9	0.7873	-0.1127	5.64	0.7917	-0.1083	7.96
	-0.5688	-0.5334	0.0354	1.53	-0.5104	0.0584	3.00
	-0.0457	-0.0393	0.0064	0.25	-0.0524	-0.0067	1.33
	-0.0647	-0.0783	-0.0136	0.67	-0.0845	0.0198	4.95
	-0.4666	-0.3289	0.1377	3.90	-0.3150	0.1516	8.13
	0.1919	0.1239	-0.0680	2.68	0.0709	-0.1210	8.78

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
I	AR	0.2914	2.98	0.3075	0.1832	1.87	0.1973
	MA	0.3675		0.3744	0.2547		0.2599
II	MA	0.1304	1.38	0.1316	0.1276	1.35	0.1288
IV	AR	0.1412		0.1807	0.0962		0.1449
	MA	0.1639		0.1677	0.1375		0.1494
	MA	0.1803		0.1804	0.0357		0.0363
	MA	0.1441		0.1447	0.0283		0.0345
	MA	0.2497		0.2852	0.1319		0.2009
	MA	0.1795		0.1919	0.0974		0.1553

Table 6.6

Summary Statistics for PH/U and PH/R

Models I, II, and IV: Sample Size N = 100

PART A

Model	P	PH/U			PH/R		
		M	B	B/SE	M	B	B/SE
I	0.9	0.8339	-0.0661	1.95	0.8551	-0.0449	2.05
	-0.7062	-0.6613	0.0449	1.04	-0.6662	0.0400	1.22
II	-0.6092	-0.6167	-0.0071	0.56	-0.6159	-0.0063	0.50
IV	0.9	0.8185	-0.0815	5.45	0.8419	-0.0581	5.92
	-0.5688	-0.5459	0.0229	1.45	-0.5443	0.0245	2.26
	-0.0457	-0.0785	-0.0328	1.91	-0.0546	-0.0089	2.45
	-0.0647	-0.0557	0.0090	0.57	-0.0728	-0.0081	2.50
	-0.4666	-0.3800	0.0866	3.55	-0.3810	0.0856	6.12
	0.1919	0.1451	-0.0468	2.74	0.1296	-0.0623	4.99

PART B

		SD	SD/ASD	RMSE	SD	SD/ASD	RMSE
I	AR	0.2399	2.93	0.2488	0.1531	1.87	0.1595
	MA	0.3060		0.3093	0.2316		0.2350
II	MA	0.0897	1.13	0.0900	0.0889	1.12	0.0891
IV	AR	0.1058		0.1336	0.0694		0.0905
	MA	0.1114		0.1137	0.0768		0.0806
	MA	0.1217		0.1260	0.0257		0.0272
	MA	0.1125		0.1129	0.0229		0.0243
	MA	0.1727		0.1932	0.0989		0.1308
	MA	0.1206		0.1294	0.0882		0.1080

Tables 6.4-6.6 bear witness to the greater efficiency of PH/R in both the sense of smaller variance and mean square error, although some of the efficiency gains (particularly for α_2 and α_3 of Model IV) must be regarded as too high. An explanation of these exists in that, for a number of replications with Model IV when Muller's method was not able to find the roots accurately, the parameters were set at the values attained before the routine broke down rather than delete the replication. Therefore the bias may be reduced from what it would be if the replications were deleted, and this unsatisfactory feature relegates Model IV to the status of a mere indicator of the feasiblity of applying the estimators.

It is apparent that both PH/R and PH/U have variances substantially above the M.L. estimator for Model I even in the largest sample, and this could be accounted for by two replications - 43 and 44 - which yielded $\hat{\alpha} > 0$ (i.e. violated the sign constraint) for PH/U. With these values deleted the mean and standard deviations of PH/U and PH/R were (0.8762, 0.1089) and (0.8835, 0.0831) respectively, which agree closely with the M.L. moments. An alternative explanation for the large biases of this model might rely on its similarity to Model A of Chapter 3.

No summary statistics are given for $\hat{\lambda}$ as the mean was generally around 10^8, and only when the sample size was 100 for Model II was the answer (5.915) sensible. An explanation of this behaviour for Model II follows and it extends (by analogy) to the other models. Consider the covariance generating function for u(t) in Model II.

i.e. $$\Gamma_{uu}(L) = 1+\lambda(1-L)(1-L^{-1}) = (1+\alpha_1 L)(1+\alpha_1 L^{-1}) \qquad (6.20)$$

(6.20) is a second degree polynomial with roots

$$\phi_1 = \frac{-(1+2\lambda)-\sqrt{(1+2\lambda)^2-4\lambda^2}}{2\lambda} \qquad \phi_2 = \frac{-(1+2\lambda)+\sqrt{(1+2\lambda)^2-4\lambda^2}}{2\lambda} .$$

Because $\lambda > 0$ and the M.A. parameter must be negative $|\phi_1| > 1$, $|\phi_2| < 1$ so that $\alpha_1 = \phi_2$.

$$\therefore \quad \alpha_1 = \frac{-(1+2\lambda)+\sqrt{(1+2\lambda)^2-4\lambda^2}}{2\lambda}$$

or $$\alpha_1 = -(1+\frac{1}{2\lambda}) + \sqrt{\frac{1}{4\lambda^2} + \frac{1}{\lambda}} \qquad (6.21).$$

For the present we seek $\lim_{\lambda\to\infty} \alpha_1$ and $\lim_{\lambda\to 0} \alpha_1$. From (6.21)

$$\lim_{\lambda\to\infty} \alpha_1 = -1 \qquad (6.22)$$

$$\lim_{\lambda\to 0} (1 + \frac{1}{2\lambda}) \to \frac{1}{2\lambda}$$

$$\lim_{\lambda\to 0} (\frac{1}{4\lambda^2} + \frac{1}{\lambda}) \to \frac{1}{4\lambda^2}$$

$$\therefore \quad \lim_{\lambda\to 0} \alpha_1 = 0 \qquad (6.23).$$

Equations (6.22) and (6.23) show that (6.20) maps the interval $[0,\infty]$ containing λ into the interval $[-1,0]$ containing α_1, so that, if at any replication a value of $\hat{\alpha}_1$ close to -1 is encountered, $\hat{\lambda}$ must be very large, and this will account for a high mean value over all replications (alternatively one might take the limit of (6.20) as $\lambda \to 0$ and $\lambda \to \infty$). Some idea of the probability of this occurrence is gained from constructing confidence intervals around the true value of α_1 with the asymptotic variance $N^{-1}(1-\alpha_1^2)$. When N = 40 the range for

$\hat{\lambda}$ will be 0.87-44, but if we adopt the empirical standard deviations of $\hat{\alpha}_1$ this range becomes 0.4-4281. All in all the above argument establishes an à fortiori case for there being a reasonable probability of very high estimates of λ in small samples, with this tendency diminishing as the sample size rises. As mentioned previously, when N = 100, the mean of $\hat{\lambda}$ was 5.915 which is quite close to the true value of 4.0. [12]

There is one other computational feature that deserves comment. This relates to the difficulty of inversions etc. associated with PH/R and PH/U and Table 6.7 contains a traditional quantity for assessing this, where replication 1 of Model I (N = 100) was selected.

Table 6.7

Condition Number and Other Features of PH/R and PH/U

Iteration Number	Estimator					
	PH/U			PH/R		
	$\hat{\beta}$	$\hat{\alpha}_1$	Cond.No.	$\hat{\beta}$	$\hat{\lambda}$	Cond.No.
1	0.9	-0.7	$0.2\text{x}10^4$	0.9	10	$0.2\text{x}10^5$
2	0.84	-0.86	$0.4\text{x}10^4$	0.84	123	$0.9\text{x}10^6$
3	0.80	-0.83	$0.4\text{x}10^4$	0.84	480	$0.2\text{x}10^8$
4				0.83	3120	$0.5\text{x}10^{10}$
5				0.82	83864	$0.7\text{x}10^{13}$

The final estimate of the M.A. parameter for PH/R was -0.82 so that both the estimates of β and α_1 were close for each estimator.

---

12 It was this behaviour which caused the breakdown of Muller's method for a number of replications of Model IV. For Models I and II however the roots were accurate even for very high λ.

Note however the condition number of the matrix. There must be some doubt about the parameters of PH/R as the condition number is so large as to render any cross product matrix and its inverse highly suspect. On the other hand PH/U does not exhibit an overly large condition number. Certainly this seems to establish a good case for imposing the restrictions directly when it is possible to do so, but unfortunately it is likely to be feasible only for Models with unit roots in B(L).

Although the Monte Carlo studies have suggested that PH/R might be used in moderate sized samples it appears that the distribution of $\hat{\lambda}$ will be non-normal unless the sample size is large. Rather than conduct an exhaustive set of experiments it was decided to press ahead to empirical work in order to determine the success or failure of the algorithm in some of the areas mentioned in 6.2.[13]

6.5 Seasonal/Trend Components

Signal extraction has been performed by economists and others for many years. Electrical engineers and acoustics experts have developed filters to extract signals from noise contaminated series, and in the same way economists and economic statisticians have dealt with the extraction of seasonal signals from the background noise of the trend and irregular components. The methods for performing such tasks have been many and varied, encompassing the classical technique of complex demodulation and the non-linear ad-hocery of the X-11, X-13 programs developed by the U.S. Census Department for the seasonal adjustment of

---

13 A further factor influencing this decision was that for 20 replications of Model IV PH/R required 30 mins. of C.P.U. time.

time series. In recent years there have been repeated suggestions that models of the various components should be built and estimated and attempts at this may be found in the work of Hannan [41] and [36]. The latter article is the stimulus to the approach of this section.

To begin the analysis it is necessary to decide on reasonable forms for each of the three components Trend, Seasonal and Irregular.

(i) Trend

In the series to be modelled later there is no cyclical element (apart from the seasonal) so that the spectrum of this component will not be peaked away from the origin. Such a shape may be achieved by the use of an A.R. with real roots, but when the series is non-stationary the two most important representations are likely to be

$$(1-L)T(t) = (1+\alpha_1 L)e(t) \qquad (6.24).$$
$$(1-L)^2T(t) = (1+\alpha_1 L+\alpha_2 L^2)e(t)$$

As Box and Tiao [12] have shown the first of these yields a series with a continuously changing level, while the second varies in both level and slope (see Box, Jenkins, and Bacon [10] for a further analysis).

(ii) Seasonal

A parametric form for the seasonal component has not been discussed in detail by any author, although [18] have paid some attention to it. In this article the authors opt for a non-stationary seasonal and, analogously to the trend, filters such as $(1-L^4)$ are chosen to reproduce a changing level in the seasonal pattern of a quarterly series while $(1-L^4)^2$ would describe both a trending level and slope. Therefore if a non-

stationary seasonal seems appropriate the two most likely forms would be

$$(1-L^4)\,S(t) = (1+\alpha_4 L^4)\,e_2(t)$$
$$(1-L^4)^2\,S(t) = (1+\alpha_4 L^4+\alpha_8 L^8)\,e_2(t) \qquad (6.25).$$

In order to fully encompass the types of seasonal patterns that may occur in data it is important to examine representations for a stationary seasonal element. Two types will be distinguished.

(A) A stable seasonal pattern. This may be efficiently removed by dummy variables.

(B) A non-constant stationary seasonal. Filters such as (6.25) would generate this if there were non-unit roots in the A.R. transfer functions i.e. $(1-\beta_4 L^4)$ in place of $(1-L^4)$.

A consideration of the spectrum of S(t) brings each of these into perspective and allows the deficiency of forms such as (6.25) to be appreciated. Firstly a stable seasonal will produce a spectrum for S(t) with spikes of infinite height at the seasonal frequencies and zeroes elsewhere, whereas all other filters tend to "spread out" about these frequencies with the degree of spread being determined by the transfer functions. Secondly (6.25) must also result in a peak at the zero frequency and in general this peak will be equal to those at the seasonals (if $\alpha_4 = 0$), thereby causing the effects of trend and seasonal to become mixed. Ideally one would like to have only peaks at the seasonals (as is the case with Hannan's models) in the spectrum of S(t).

(iii) Irregular

A convenient assumption is that the irregular is white noise but this may be misleading in that the effects of strikes and working day variations may induce some correlation. For most of these special factors dummy variables would be appropriate and in what follows only the simplest assumption is adopted

$$I(t) = e_3(t) \qquad (6.26).$$

The first series to be examined will be the wheat yields data analysed by Schmitz and Watts [98]. These authors fitted a number of ARMA models and concluded that a first order M.A. fitted to first-differenced data would yield a satisfactory explanation. As the data was annual - and there were no peaks in the spectra at any frequency - only two components will be considered, and to reflect the Schmitz-Watts conclusion regarding the type of model to be employed these will be

$$\begin{aligned} T(t)(1-L) &= e_1(t) \\ I(t) &= e_2(t) \end{aligned} \qquad (6.27).$$

Schmitz and Watts fitted the reduced form of(6.27) by PH/O and Table 6.8 presents the parameter estimates for Model II when PH/R is used. Only two countries - Canada and the United States - will be dealt with.

Table 6.8

Model II fitted to Canadian and United States Wheat Yields

1900-1960

Country	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\sigma}^2$
Canada	-0.7616	13.41 (1.35)	21.24
United States	-0.4501	1.49 (1.04)	3.56

The main item of interest in Table 6.8 is the insignificance of $\hat{\lambda}$ for both countries. This was usual for all models fitted and seems to indicate that the numerical covariance matrix may be misleading.[14]

We now turn to the extraction of the trend from the wheat yield data. This is accomplished with the aid of (6.6) and Table 6.9 gives the trend extraction and the original series for Canada and the United States for the period 1940-1960.[15]

---

14 Alternatively one might fit the models with PH/U. When this was done the estimates for both were very close and a t-test performed on $\hat{\alpha}_1$ revealed that it was significantly different from zero. As shown previously $\lim_{\lambda \to 0} \alpha_1 = 0$ so that there is a discrepancy in the two conclusions.

15 There is a "start up" problem in using (6.6) i.e. what is the value of $\hat{T}(-1)$? Two solutions were proposed. Firstly set $\hat{T}(-1) = 0$. This may not be satisfactory for the Canadian case in which the root of the M.A. is quite high, but should be reasonable for the U.S.. Secondly set $\hat{T}(-1) = y(0)$. The discrepancy should be much smaller under this alternative and therefore the estimated series $\hat{T}(t)$ would be more reliable. In practice, for the period 1940-1960, both suggestions gave almost the same answers.

Table 6.9

Estimates of the Trend Component in the Wheat Yields Data

Canada and the United States : 1940-1960

Year	Canada		United States	
	y(t)	$\hat{T}(t)$	y(t)	$\hat{T}(t)$
1940	18.80	14.84	15.30	14.58
1941	14.40	14.73	16.80	15.80
1942	25.80	17.37	19.50	17.84
1943	16.90	17.26	16.40	17.05
1944	17.90	17.41	17.70	17.41
1945	13.60	16.50	17.00	17.18
1946	16.90	16.60	17.20	17.19
1947	14.10	16.00	18.20	17.75
1948	16.00	16.00	17.90	17.83
1949	13.50	15.41	14.40	15.94
1950	17.10	15.81	16.60	16.30
1951	21.90	17.26	16.00	16.14
1952	26.50	19.46	18.30	17.33
1953	24.10	20.57	17.30	17.31
1954	13.70	18.93	20.40	19.01
1955	24.10	20.16	20.90	20.05
1956	25.20	21.36	20.80	20.46
1957	17.60	20.47	22.50	21.58
1958	17.80	19.83	28.60	25.44
1959	17.90	19.37	23.20	24.21
1960	21.20	19.81	27.80	26.18

Once a series $\hat{T}(t)$ has been constructed (6.6) allows the decomposition of any new observation into trend and irregular factors. As an example take the 1961 observation for Canada i.e. y(1961) = 11.20. Applying (6.6) with $\hat{T}(1960)$ = 19.81 (from Table 6.9), $\hat{T}(1961)$ = 17.76 so that the irregular was 6.56.

Two quarterly time series will now be dissected. One of these will be New Money which exhibited unit roots in the M.A. when the various estimators of Chapter 4 were applied and some reasons will now be advanced for that behaviour. Fig. 6.1 graphs this series and leaves the suspicion that there may well be a stationary seasonal pattern. To test various hypotheses concerning the trend and seasonal the following models-embodying these possibilities - were fitted, and the resulting equations were

$$\underset{\quad(2.94)\ (5.21)\ (1.78)\ \ (2.45)\ \ (0.61)}{y(t) = 35.77 + 1.34t + 22.56\,S_1 + 30.92\,S_2 - 7.68\,S_3} \qquad \sigma^2 = 1193.7,\ R^2 = 0.41 \qquad (6.28)$$

$$\underset{(12.39)\ \ (1.62)\ \ (3.10)\ \ (8.38)\ \ (2.46)\ \ (0.88)}{(1-0.8L)y(t) = -13.56 + 24.23\,S_1 + 64.08\,S_2 + 19.48\,S_3 + (1-0.14L)e(t)} \qquad (6.29)$$

$$\sigma^2 = 392.1$$

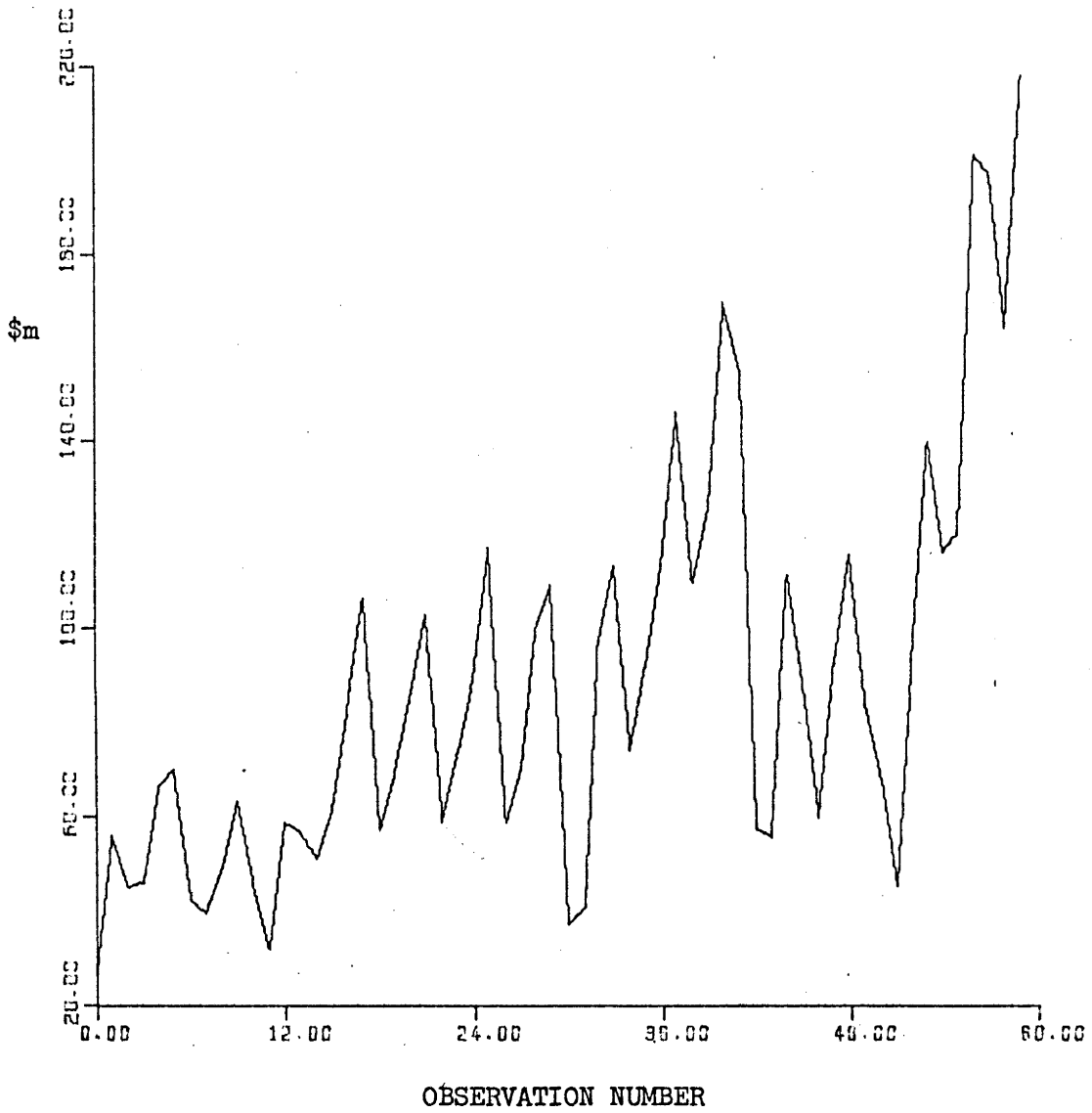
$$\underset{(7.21)\ \ (0.61)}{(1-L)(1-0.70L^4)y(t) = 0.57 + (1-0.29L-0.45L^2-0.10L^3-0.36L^4+0.003L^5)e(t)} \qquad (6.30)$$

$$\sigma^2 = 588.3$$

In (6.28) and (6.29) S_i (i=1,2,3) are dummy variables with values 1.0 in financial quarter i and zero otherwise, while t is a time trend. (6.30) is derived from a three components model and $\hat{\lambda}_1$ was 124.96 while $\hat{\lambda}_2$ was 56.18. Of the three models there was extensive autocorrelation in the residuals correlogram for (6.28) with the ordinates declining as an A.R., while for (6.29) the highest t statistic was 0.81 at lag 6. Comparing the residual variances to Table 4.24 there is no doubt that the second of these models provides by far the best fit of any suggested, and this leads to the conclusion that the series may well

FIGURE 6.1

NEW MONEY



have a stationary trend component and (probably) a stable seasonal as well (although a combination of a stationary but non-constant seasonal and a stationary trend might also fit well). The precise parametric form is not important at this juncture and it will serve merely to notice that at most only one of the components will be non-stationary. What will be the impact of differencing? The answer is obviously the introduction of unit roots into the M.A., and this is what was found in Chapter 4.[16] This example should serve as a warning against a too hasty use of differencing. When the data has only two components a visual inspection will usually serve to indicate if non-stationarity is present, but when there is a seasonal pattern it may be difficult to tell, and this highlights the need for some non-parametric tests for the presence of certain types of evolving behaviour in time series.[17]

Finally some experiments were carried out on the Miscellaneous consumption series of Chapter 5. It is assumed that the three components may be modelled by

---

16 This is very easily seen in a two components model. Assume that the trend is described by $T(t) = a + bt$ and the irregular by $I(t) = e_2(t)$. Then differencing $y(t)$ will yield

$$(1-L)y(t) = a + bt - a - b(t-1) + (1-L)e_2(t)$$

$$\therefore (1-L)y(t) = b + (1-L)e_2(t).$$

Similarly fourth differencing of a set of dummy variables would produce the same effect for the seasonal.

17 One that springs to mind is the classical technique of complex demodulation. To test for a non-stationary seasonal, demodulation would be performed at the seasonal frequencies, the series re-modulated, and tests constructed for the presence of evolutionary behaviour. An attempt was made along these lines for New Money but the filter required to effectively extract the demodulate resulted in a severe loss of observations, and for this reason the technique may not be applicable to quarterly time series. Nevertheless it is worth reporting that the hypothesis of a stationary seasonal pattern was supported.

$$(1-L)T(t) = e_1(t)$$
$$(1-\beta_4 L^4)S(t) = e_2(t)$$
$$I(t) = e_3(t).$$

The reduced form (fitted by PH/R) was

$$(1-L)(1-0.93L^4)y(t) = (1-0.19L-0.93L^4 + 0.18L^5)\,\varepsilon(t) \quad \sigma^2 = 352.7(6.31),$$

where α_2 and α_3 were both less than 10^{-3}. The fit of this is slightly superior to any examined in Section 5.5, and, as it does not suffer from the unit root problem encountered there, shows that components models are desirable for their ability to produce a good fit to the data quite apart from any signal extraction objective. Although it was not done, both the trend and seasonal elements in Miscellaneous could be extracted in a similar fashion to the earlier exercise with wheat yields. The procedure is much more complex though in that there must now be a partial fractions expansion of an eleventh order polynomial.

6.6 A.R. with superimposed noise

The lure of a possible A.R. with superimposed noise (A.R.S.N.) in the sunspots data was impossible to resist. Accordingly the series in [4 p.660] was selected and this provided 176 observations covering the period 1749-1924 (inclusive). Initially it was assumed that the A.R. was of second order and the resulting parameter estimates are contained in Table 6.10.

Table 6.10

Parameter Estimates for the Sunspots Data

A.R. (2) with superimposed noise*

Estimator	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\lambda}$	$\hat{\sigma}^2$
O.L.S.	13.9	1.34	-0.65				241.6
	(7.0)	(23.0)	(11.2)				
PH/U	13.7	1.40	-0.71	-0.11	0.03		240.0
	(5.7)	(13.5)	(8.8)	(0.8)	(0.3)		
PH/R	13.8	1.40	-0.71	-0.11	0.03	0.06	238.8
	(7.5)	(20.3)	(10.6)			(1.0)	

*$\hat{\mu}$ refers to the estimated mean; all other symbols should be familiar.

Table 6.10 supports the rejection of a second order A.R.S.N. model as the mechanism generating the sunspots data. Neither the M.A. parameters of PH/U nor $\hat{\lambda}$ of PH/R are significantly different from zero-implying the absence of a second component. A comparison of the sum of squares (remembering that PH/U estimates 4, and PH/R 3 more parameters than O.L.S.) also sustains the null hypothesis. Finally the correlogram of residuals from the O.L.S. regression-presented in Table 6.11 - reveals that the disturbance is close to white noise.

Table 6.11

Correlogram of the Residuals from O.L.S.

Untransformed Sunspots Data

Lag	$\hat{\rho}$	t
1	-0.07	0.91
2	0.07	0.97
3	-0.05	0.69
4	0.10	1.37
5	-0.02	0.31
6	0.04	0.50
7	-0.08	1.01
8	0.08	1.07
9	0.13	1.69
10	0.09	1.13
11	0.16	2.11
12	0.11	1.46
13	0.01	0.08
14	-0.03	0.37
15	0.00	0.02
16	-0.08	1.12
17	0.09	1.14

The apparent absence of an A.R.S.N. model in the untransformed data led to the decision to follow Bailey and smooth the data by taking the logarithm of each observation after 6.0 had been added. Having done this he argued that a fifth order A.R. was necessary to describe the new data set i.e. in the context of equation (6.1) p would be 5. Table 6.12 lists the parameter estimates associated with the three estimators (where e* was set at zero).

Table 6.12

Parameter Estimates for the Transformed Sunspots Data

A.R. (5) with superimposed noise

Parameter	Estimator		
	O.L.S.	PH/U	PH/R
$\hat{\mu}$	1.43 (6.62)	0.41 (4.72)	0.42 (3.55)
$\hat{\beta}_1$	1.23 (15.71)	0.77 (5.16)	2.60 (16.10)
$\hat{\beta}_2$	-0.48 (3.87)	0.78 (4.42)	-3.43 (8.49)
$\hat{\beta}_3$	-0.10 (0.77)	-0.89 (5.71)	2.89 (6.07)
$\hat{\beta}_4$	0.05 (0.36)	-0.40 (2.13)	-1.62 (5.25)
$\hat{\beta}_5$	-0.08 (0.99)	0.37 (2.88)	0.45 (4.36)
$\hat{\alpha}_1$		1.36 (3.15)	-1.43
$\hat{\alpha}_2$		0.46 (3.17)	1.46
$\hat{\alpha}_3$		-0.77 (5.30)	-1.00
$\hat{\alpha}_4$		-0.36 (2.48)	0.43
$\hat{\alpha}_5$		0.50 (4.59)	-0.09
$\hat{\lambda}$			1.48 (2.20)
$\hat{\sigma}^2$	0.1302	0.1189	0.1279

There is wide variation in the parameter estimates but not a great difference in the residual variances. A likelihood ratio test performed on each estimator for equality of variances yielded the following results:-

(i) O.L.S. and PH/R were not significantly different (at the 5% level).

(ii) O.L.S. and PH/U were significantly different.

(iii) PH/R and PH/U were also significantly different.

From these comparisons it appears that the data will not support the restrictions implied by an A.R.S.N. model (as embodied in PH/R) and that the best fit would come from a general ARMA model. An appreciation of the reasons for the failure of PH/R may be gained by referring to the correlogram of O.L.S. residuals presented in Table 6.13.

Table 6.13

Correlogram of Residuals from O.L.S.

Transformed Sunspots Data

Lag	$\hat{\rho}$	t	Lag	$\hat{\rho}$	t
1	0.01	0.09	11	0.15	1.97
2	0.02	0.21	12	0.22	2.90
3	0.04	0.49	13	-0.01	0.09
4	0.09	1.23	14	-0.03	0.42
5	0.09	1.20	15	0.04	0.51
6	0.03	0.45	16	0.01	0.13
7	0.12	1.51	17	0.04	0.52
8	-0.08	0.98	18	-0.08	1.01
9	0.19	2.53	19	0.01	0.10
10	0.11	1.43	20	-0.08	1.01

Table 6.13 indicates that the poor performance of PH/R may be explained by the absence of a fifth order M.A. in the residuals which should be found if an A.R.S.N. model were appropriate.[19] Therefore the particular version of A.R.S.N. employed is probably incorrect.

---

[19] One should be wary of this interpretation because O.L.S. will not be consistent if a fifth order M.A. were present in the data and the estimated correlogram estimates would also be inconsistent. It is interesting to note that [9 p.186] identify the appropriate model for the sunspots data as a second order A.R.

Finally Table 6.14 compares the roots of $A(L^{-1})$ from PH/R, PH/U and Bailey's article.

Table 6.14

Roots of $A(L^{-1})$ for the Transformed Sunspots Data

Bailey	PH/R	PH/U
0.22	0.49	-1.04
0.40 ± 0.32i	0.41 ± 0.40i	-0.56 ± 0.67i
0.46 ± 0.04i	0.06 ± 0.75i	1.42 ± 0.76i

It is interesting to note that only two of Bailey's estimated roots are close to the M.L. values, and this leads to the conclusion that the lack of an iterative refinement in his estimator does not enable an ordered approach to the M.L. estimates. Such an admission must severely circumscribe the practical uses of his estimator.

6.7 Permanent Income

This was one of the examples of Section 6.2 wherein a signal had to be extracted from a noisy background and the mode of analysis will be similar to that for the wheat yields data. As mentioned earlier the A.R. parameter selected by Friedman implies a non-stationary series for income so that the first step was to examine the correlogram of Personal Disposable Income 1905-1949 (1929 prices).[20] Table 6.15 presents the first ten lags.

---

[20] Taken from Goldsmith [31 Vol.III]. Friedman excluded the war years 1917, 1918, 1942 through 1945 in his analysis, but the conclusions were not qualitatively altered by working with the complete series.

Table 6.15

Correlogram of Personal Disposable Income 1905-1949 (1929 Prices)

Lag	$\hat{\rho}$	t
1	0.91	6.11
2	0.82	5.49
3	0.73	4.89
4	0.63	4.20
5	0.55	3.68
6	0.46	3.11
7	0.38	2.56
8	0.30	2.00
9	0.22	1.48
10	0.17	1.13

Table 6.15 shows that the correlogram is reminiscent of a first order A.R. with $\beta \simeq 0.9$ and it seems certain that the series is stationary. Why then did Friedman select a value for β characteristic of a non-stationary series? Regarding this choice Friedman argues that unless $\beta > 1$ there is no allowance for secular growth i.e. "Being an average of earlier observations the estimated y_p^* is necessarily between the lowest and the highest, so that this method of estimation applied to a steadily growing series yields estimated values systematically below the observed values" [28 p.144]. This argument is not convincing as surely the essence of permanent income is that it lag behind changes in actual income. Nevertheless $\beta = 1$ might be realistic.

Although it is obvious that the signal extraction approach to the construction of a permanent income series cannot be identical to Friedman's (unless $\hat{\beta} = 1.\dot{3}$ and there must be serious doubts about the properties of any ARMA estimator under such conditions) we will attempt to apply the methodology given earlier. Owing to the pattern exhibited

by the correlogram the assumed components model took the form

$$\begin{aligned}(1-\beta L)\,(T(t)-\mu) &= e_1(t)\\ I(t) &= e_2(t)\end{aligned} \qquad (6.32),$$

where T(t) stands for permanent, and I(t) for transitory income. The estimated reduced form was

$$\begin{aligned}y(t)(1-0.9379L) &= 72683 + (1-0.00002L)\,e(t)\\ (26.86) \qquad\qquad & \qquad\qquad \sigma^2 = 277.9\end{aligned} \qquad (6.33).$$

In (6.33) $\hat{\lambda} = 625281$ with a t value of 0.47,[21] meaning that the variance of the transitory component is very small when compared to the variance of the disturbance entering the permanent element. By substituting $\hat{\beta},\hat{\alpha}$ into (6.6) it is seen that $\hat{T}(t) \simeq y(t)$ i.e. there is no transitory income in the series. Although such a conclusion offends intuition it is an inescapable concomitant of the close fit achieved by a pure A.R.. Therefore this foray cannot be counted a great success. However it may be that the level of aggregation implicit in the income series makes nonsense of concepts based upon the rationalization of an individual entity and the difficulty might be overcome with disaggregated data.

6.8 Conclusion

The sections of this chapter have ranged over a wide area in which signal/noise models are to be found. There was a very cursory discussion on the sampling properties of the M.L. estimator PH/R, but even from this there was established an awareness that very large samples might be needed for its successful employment. Rather than pursue this line of enquiry an attempt was made to assess the utility of the estimator (and underlying model) by analysing some of the applications

[21] $\hat{\lambda} = \frac{1}{\hat{\lambda}}$ i.e. $\hat{\lambda}$ is the ratio σ_1^2/σ_2^2.

ascribed to it in Section 6.2. Generally the conclusion must be negative, but whether this was a fault of insufficient experimentation or the basic model is unclear. Certainly the evidence presented concerning the presence of superimposed noise in the sunspots data is at variance with the ideas of many statisticians in past years, and it may be that the technique derives value from such unexpected results.

Appendix 10

Computational Difficulties with the Estimator PH/R

As mentioned in the text, finding the roots of $\Gamma_{uu}(L)$ proved to be troublesome. At first there was the complete breakdown of many standard programs for solving polynomials and later, even when Muller's method was in operation, cases were discovered where the roots were not determined to the desired degree of accuracy. This arose from the rounding errors associated with the synthetic division of the polynomial reduction process (see Ralston[95 p.371]). After all divisions are complete the remainder should be 1.0 (if there is no rounding error) but with single precision arithmetic i.e. accuracy to around 6 decimal places, the remainder was generally of the order of $L + 10^7$. Implementing a double precision version reduced the remainder to $L + 0.5$ thereby ameliorating the problem without solving it.

After some experimentation it emerged that this feature was due to the inverse relation between the roots e.g. if 0.02 was a root then so was 50. As a result of the large differences between the coefficients attached to each power of the quadratic formed from these roots e.g. $L^2 + 50.02L + 1$ synthetic division produced a reduced polynomial with incorrect parameters. Therefore it seemed desirable to find roots close to unity first and then to work down toward zero. In Muller's method (as presented in [17]) the first three root approximations are (-1,1,0), causing a tendency to initially seek a root close to zero. By the simple manoeuvre of setting the maiden guesses at (0,2,1) the positive root closest to unity was always extracted before any others. No theoretical justification can be given for this but the evidence of its effectiveness can be seen from the fact that in most instances the remainder was $1 + 10^{-14}$ - the greatest accuracy obtainable with the

360/50. However even after these modifications there were still cases when the roots of $\Gamma_{uu}(L)$ were incorrectly determined e.g. if the variance ratio became very large, and future investigators might experiment with such methods as Lehmer-Schur [95] and/or arithmetic of greater precision than that available as a standard option on most computers.

Another finding was that large values of λ were handled more easily than small because of the boundary restriction $\lambda > 0$. By choosing a suitable normalization rule computer time was not wasted in slow convergence near to a boundary.

Appendix 11

Derivation of Bailey's Estimating Equation

The model to be investigated is

$$y(t) = B^{-1}(L)\, e_1(t) + e_2(t) \qquad \text{(A11.1)},$$

which becomes (in Fourier notation)

$$Y(\lambda) = B^{-1}(\lambda)\, e_1(\lambda) + e_2(\lambda) \qquad \text{(A11.2)}.$$

Our aim will be to find an optimal predictor of the form

$$y^*(t) = W(L)\, y(t) \qquad \text{(A11.3)},$$

or

$$Y^*(\lambda) = W(\lambda)\, Y(\lambda) \qquad \text{(A11.4)}.$$

To do this we minimize the prediction error variance

$$V = E\left[\frac{1}{2\pi N}\,(Y(\lambda) - Y^*(\lambda))^2\right] \qquad \text{(A11.5)}$$

with respect to the elements of W(L).

Substituting (A11.2) and (A11.4) into (A11.5) and re-arranging the result leaves

$$V = E\left[\frac{1}{2\pi N}\,\{B^{-1}(\lambda)\, e_1(\lambda) + e_2(\lambda)\}\{1-W(\lambda)\}\right]^2 \qquad \text{(A11.6)}.$$

Expanding (A11.6) we obtain

$$V = E[\{|B(\lambda)|^{-2} I_{11}(\lambda) + B^{-1}(\lambda) I_{21}(\lambda) + \overline{B}^{-1}(\lambda) I_{12}(\lambda) + I_{22}(\lambda)\}$$
$$\{1 - W(\lambda) - \overline{W}(\lambda) + |W(\lambda)|^2\}] \qquad \text{(A11.7)},$$

where $I_{k\ell}(\lambda)$ is the cross periodogram between $e_k(t)$ and $e_\ell(t)$.

Taking the expectation of the expression in (A11.7) and simplifying, V becomes

$$V = [(2\pi)^{-1}\, \sigma_1^2 |B(\lambda)|^{-2} + (2\pi)^{-1}\, \sigma_2^2][1 - W(\lambda) - \overline{W}(\lambda) + |W(\lambda)|^2] \qquad \text{(A11.8)}.$$

Differentiating (A11.8) with respect to the k'th element of W(L) (k=0,...,K) and setting $\frac{\partial V}{\partial \omega_k} = 0$ yields

$$[(2\pi)^{-1}\sigma_1^2|B(\lambda)|^{-2}+(2\pi)^{-1}\sigma_2^2][-e^{i\lambda k}-e^{-i\lambda k}+W(\lambda)e^{-i\lambda k}+\bar{W}(\lambda)e^{i\lambda k}] = 0$$

$$k = 0,\ldots,K \qquad (A11.9)$$

or multiplying through by $|B(\lambda)|^2$

$$[(2\pi)^{-1}\sigma_1^2+(2\pi)^{-1}\sigma_2^2|B(\lambda)|^2][-e^{i\lambda k}-e^{-i\lambda k}+W(\lambda)e^{-i\lambda k}+\bar{W}(\lambda)e^{i\lambda k}] = 0$$

$$k = 0,\ldots,K \qquad (A11.10).$$

(A11.10) is a system of equations linear in ω_k that may be put in the form of (6.17) in the text.

Bailey's approach differs from the above in two minor ways:

(i) It is formulated in the time domain with the aid of lag operators. The transition to this form is easily made by the substitution $L=e^{i\lambda}$.

(ii) Bailey does not assume that $E(e_1(t)\, e_2(t)) = 0$ i.e. $E(I_{12}(\lambda)) \neq 0$, but to preserve the standard model this restriction was employed in deriving (A11.8).

CHAPTER 7

The Utility of ARMA/ARMAX Models

7.1 Introduction

In this final chapter some of the threads of previous chapters are gathered and woven together as a guide to future research. This will be facilitated by a study of the utility of the ARMA/ARMAX forms in econometrics, both as it is expressed in published papers to date and may emerge in the future. By such an investigation it is to be hoped that, as well as a demonstration of the range of application of the techniques discussed in earlier chapters, a better appreciation of the ways in which these models are of importance to econometricians will emerge.

7.2 The Utility of ARMA Models

Four main uses may be distinguished for ARMA models.

(i) Forecasting tools

This was the original justification proposed by Box and Jenkins who demonstrated the power of ARMA models in forecasting time series such as I.B.M. share prices and monthly airline passenger totals. Since their pioneering work there have been a number of other studies relying on the ARMA form e.g. Struik [102] and Leuthold et al [66]. Struik reports on the success of ARMA models in forecasting a number of Australian economic time series, while Leuthold et al. compare forecasts of pig prices and quantities from this time series formulation with those originating from estimated demand and supply equations of the cobweb variety. Although the "full" econometric model gave superior forecasts the differences were not great (the Theil inequality coefficients being 0.70 and 0.65) so that, bearing in mind the smaller information requirements for ARMA equations, this alternative may provide

forecasts at smaller expense both in terms of time spent on the exercise and the computer. Obviously this is one of the areas of applied research that should be studied in greater detail and many econometric formulations that have tended to become conventional wisdom e.g. Jorgenson's neo-classical investment equation, should have their forecasts compared with those from an appropriate ARMA model.

(ii) Optimal Extractions

Chapter 6 has shown how ARMA equations may be exploited to construct estimates of trend components, permanent income etc.. There is one other possibility that was not given attention there and this relates to the formulation of expectations. In Chapter 1 two commonly used hypotheses concerning the formation of these were shown to result in ARMA models, and this leads one to conjecture that the parameters of such "expectation-generating mechanisms" might be identified with the parameters of ARMA equations. The methodology of Chapter 6 may be applied here. If there is an internal structure to the price series - in the sense that it may be viewed as being composed of a normal level with superimposed random variations - then we once again have a "signal plus noise" model. Restricting the evolutionary behaviour of the normal price to a first order A.R. enables the optimal predictor of the next period price to be derived as an exponential weighting of all past prices i.e. it coincides with the adaptive expectations hypothesis. Varying hypotheses will imply different types of ARMA equations for the price series so that it is possible to test these by seeing if the underlying model is appropriate.[1]

[1] It is worthwhile noting that the random walk model found for stock (and pig [66]) prices would be compatible with adaptive expectations only if $\alpha_1 < 0$

(iii) Naive Models

One of the tests of specification of an econometric model usually resorted to is the extent of its predictive superiority - within and subsequent to the sample period - over so called "naive" models. Unfortunately there is no theory available for guiding the selection of the latter and in many cases the ad-hoc formulations adopted provide no real test of the alternative.[2] In many cases A.R.'s have been proposed as suitable candidates for the role but from both a theoretical and empirical viewpoint it is obvious that ARMA models must be a better choice. Once this is recognized it merely remains to select the best fitting model of this class by the process of specification, estimation, and diagnostics proposed in [9]. Nelson [79] has made a promising beginning along these lines in comparing the F.R.B. - M.I.T. - Penn. model forecasts to those generated by various ARMA models: although the comparison may have been biassed against the naive models in that no account was taken of the considerable correlation between the residuals of some equations. Nevertheless his findings were revealing in that, although the econometric model forecasts were preferable to the ARMA predictions within the sample period, the rankings were reversed when outside the period.[3] Considering the expense involved in building a full-scale econometric model there are likely to be major benefits stemming from further inquiries of this type.

---

2 An example of this is the use in [98] of an exponential smoothing formula (fitted by discounted least squares) as a control solution against which ARMA forecasts might be judged. Unless one is prepared to give the "naive" version a chance the exercise is pointless.

3 The exogenous variables used by Nelson were the true value so that a source of error in the F.R.B.-M.I.T.-Penn. model has been omitted.

(iv) Estimated Exogenous Variables

One of the difficulties sometimes encountered in forecasting with small-scale econometric models is the construction of series of future values of the exogenous variables. Because ARMA models only require knowledge of the history of the exogenous variable itself they may be ideally suited to this task. Certainly it would be a valuable exercise to combine these forecasts with an econometric model and compare the resulting endogenous variable predictions with those predicated on the true values.

7.3 The Utility of ARMAX Models

7.3.1 The Structure of the Error

Chapter 1 has outlined the Thomas and Wallis challenge to the assumption of first order autocorrelation in the residuals when quarterly data are to be manipulated, but, even if the presence of high order autocorrelation is conceded, there is likely to be some dispute over a suitable parametric form for it. For many years economists have dwelt upon the first order A.R., and the influence of this tradition may be seen in the Thomas and Wallis specification of an A.R. - albeit fourth order. Yet these schemes exhibit distinctive correlograms e.g. with a fourth order A.R. one should observe steadily declining peaks in the correlogram at the fourth, eighth, etc. lags, so that an inspection of the residual autocorrelation function may be useful in deciding on a model to adequately describe the disturbances. It is this aspect that will be examined in the current section.

Two equations will provide the basic building blocks. These come from the Reserve Bank of Australia's (R.B.A.) monographs [86][87]

and refer to Personal Consumption Expenditure on Other Durables and Employment respectively. For the former a stock adjustment model of the Stone-Rowe type was found to provide a good fit to the data, the reduced form being

$$C_t = b_1+b_2 D_1+b_3 D_2 + b_4 YPD_t + b_5 KOD_{t-1} + b_6 POD_t+b_7 NC_t+u_t \quad (7.1)$$

where C_t Real Consumption Expenditure on Other Durables

YPD_t Real Personal Disposable Income

KOD_{t-1} Stock of Other Durables lagged one quarter

POD_t Price of Other Durables deflated by a general price index

NC_t A variable representing either New Instalment Credit for Retail Sales or New Instalment Credit plus changes in Bank Advances to persons.

D_i $(i=3,4)$ Seasonal dummies with value 1 in calendar quarter i and zero otherwise.[4]

The available data constituted a sample of 36 observations and (7.1) was fitted. With the two alternative definitions of NC_t the fitted equations correspond to (C.7) and (C.8) of [86] and will be referred to as (7.1a) and(7.1b) in future.

It will be necessary to present a formal derivation of the employment equation with a view to later modifications. Underlying the R.B.A. formulation is a neo-classical theory of factor demand and a stock adjustment model. To give operational meaning to the general theory it is necessary to specify a form for the production function and

---

[4] The other seasonal dummy was insignificant and was therefore omitted by R.B.A.

it will be assumed that the Constant Elasticity of Substitution (C.E.S.)[5] version links Output (Y_t), Capital (K_t) and Labour (L_t)

$$Y_t = \alpha[\ K^{-\gamma} + (1-\delta)\ L^{-\gamma}]^{-\frac{1}{\gamma}} \qquad (7.2),$$

Under the assumptions of

(a) Perfect competition in both factor and commodity markets.

(b) Profit maximizing behaviour by all production units, the marginal revenue product of labour will be equated to its price (w_t)

$$\text{i.e. } L_t^* = A \left(\frac{p_t}{w_t} \right)^{\beta} Y_t \qquad (7.3)$$

where L_t^* is the desired stock of labour

p_t is the price of output Y_t

$\beta = 1/1+\gamma$

$A = (1-\delta)^{\beta}\ \alpha^{-\gamma\beta}$.

(7.3) corresponds to (2) in [87] if the non-linear restrictions between A and β are ignored (and there would seem to be little lost by so doing).

The stock adjustment enters as a consequence of the immobility of labour in the short-run : changes in desired employment demand are only manifest in actual labour flows with a time lag

$$\text{i.e. } L_t = W(L)\ L_t^* \qquad (7.4).$$

Substituting (7.3) into (7.4) the simplified expression is

$$L_t = H(L)\ [(\tfrac{p}{w})^{\beta} Y\ ]_t \qquad (7.5),$$

---

5 This differs from the R.B.A.'s use of the Cobb Douglas production function but as will be seen the C.E.S. underlies their regression equation. Although not recognizing this explicitly they concede that adoption of their final equation "..... amounts to rejecting the Cobb Douglas production function" [87 p.6].

or under a rational lag specification $H(L) = \frac{U(L)}{V(L)}$

$$V(L)L_t = U(L)\left[\left(\frac{p}{w}\right)^{\beta} Y\right]_t \qquad (7.6).$$

(7.6) is (6) of [87] and it forms the basis for the actual estimating equations (where DE_t is a dummy variable to take account of the introduction of a new employment series in 1966).

$$L_t = c_1 + c_2 D_1 + c_3 D_2 + c_4 D_3 + c_5\left[\left(\frac{p}{w}\right)^{\beta} Y\right]_t + c_6\left[\left(\frac{p}{w}\right)^{\beta} Y\right]_{t-1} + c_7 DE_t + c_8 L_{t-1} + u_t \qquad (7.7a)$$

$$L_t = c_1 + c_2 D_1 + c_3 D_2 + c_4 D_3 + c_5\left[\left(\frac{p}{w}\right)^{\beta} Y\right]_t + c_6\left[\left(\frac{p}{w}\right)^{\beta} Y\right]_{t-1} + c_7 DE_t + c_8 L_{t-1} + c_9 L_{t-2} + u_t \qquad (7.7b)$$

Owing to the multicollinearity in the data the order of V(L) could not be raised above 2 so that (7.7a) and (7.7b) cover the only two possibilities tested. Additionally, to avoid non-linear estimation problems, the R.B.A. performed a search over the grid $0 \leq \beta \leq 1$ for (7.1a) and found that $\beta = 0.45$. To preserve compatibility with their results this restriction will be retained, although the non-linear restriction could easily have been imposed by the program. Finally (7.7a) and (7.7b) correspond to (9) and (10) of [87] and there were 42 observations.

Equations (7.1a), (7.1b), (7.7a) and (7.7b) were estimated by O.L.S. and the correlograms of the residuals from the regressions are in Table 7.1.

Table 7.1

O.L.S. Residuals Correlogram (8 lags)

Equations (7.1a),(7.1b),(7.7a) and (7.7b)

Eq.	1	2	3	4	5	6	7	8
(7.1a)	0.06	-0.17	-0.20	0.42	-0.08	-0.06	-0.02	-0.02
	(0.38)	(1.01)	(1.23)	(2.49)	(0.50)	(0.33)	(0.09)	(1.42)
(7.1b)	0.06	-0.22	-0.19	0.40	-0.08	-0.09	-0.02	0.23
	(0.35)	(1.34)	(1.16)	(2.41)	(0.48)	(0.56)	(0.10)	(1.38)
(7.7a)	0.19	-0.30	-0.33	-0.01	-0.14	0.01	0.23	0.10
	(1.19)	(1.90)	(2.12)	(0.03)	(0.89)	(0.07)	(1.48)	(0.61)
(7.7b)	0.13	-0.26	-0.35	0.01	-0.14	0.04	0.23	0.06
	(0.81)	(1.67)	(2.24)	(0.03)	(0.88)	(0.22)	(1.43)	(0.37)

Table 7.1 suggests that there is autocorrelation present in the residuals. Although the standard errors utilized in constructing the t-tests refer to the asymptotic distribution of the serial correlation coefficients constructed from data that is not residuals from an A.R., it seems as if the test will err on the side of high Type II errors [11]. This is almost certain to be so for (7.7a) and (7.7b) as the asymptotic variance of $\hat{\rho}(1)$ would be $N^{-1}\beta^2$ (where β is the A.R. parameter) rather than N^{-1}. There are further difficulties in deciding on a plausible representation for the autocorrelation but there are two important contenders for (7.1a) and (7.1b) viz.

$$u(t) = (1 + \alpha_4 L^4)\varepsilon(t) \qquad (7.8)$$

$$(1-\alpha_4 L^4)u(t) = \varepsilon(t) \qquad (7.9)$$

i.e. a fourth order M.A. or A.R. in the disturbances. If (7.8) was satisfactory there should be a peak only at the fourth lag in the correlogram while (7.9) would have peaks at 4, 8 etc. Bearing in mind the comments concerning Type II errors it is fairly obvious that Table

7.1 does not allow discrimination between the two hypotheses so that both alternatives (along with O.L.S.) were fitted and the parameter estimates are in Table 7.2.[6]

The idea that there was fourth order residual autocorrelation is fully supported by Table 7.2, and from a comparison of the residual variances it appears to be of the A.R. variety. As both specifications resulted in the boundary restriction $\alpha_4 < 1$ being applied there may be some non-stationarity in the seasonal pattern of the residuals, which dummy variables were not able to account for. There are some interesting changes in the parameter estimates, including an increasingly significant influence for the credit variable and a reduction in the impact of the capital stock, and there now is a preference for (7.1a) over (7.1b).

Turning to the employment example, on the basis of Table 7.1 it was decided to append a third order M.A.

$$u(t) = (1+\alpha_2 L^2+\alpha_3 L^3)\varepsilon(t) \qquad (7.10)$$

to each equation. Table 7.3 contains the parameter estimates for O.L.S. and the ARMAX estimator.

---

6 It was impossible to exactly reproduce the R.B.A.'s estimates for any equation. A number of regression routines were applied with all results being similar to Table 7.2 e.g. using the simple precision I.B.M. multiple regression package (comprising CORRE,ORDER,MINV and MULTR of [54]) the parameter estimates were

(532, -5, -5, 0.0732, -0.030, -476, 0.2353) for (7.1b)

and (332, 49.1, 26.3, 13.5, 0.1658, 0.0656, 39.5, 0.5953) for (7.7a), which differ only in very minor ways from Tables 7.2 and 7.3 (p 286). The data utilized in the regressions is listed in Appendix 12. For the moment it is only necessary to note that in as much as the R.B.A. seasonal dummy estimates were the same as those in Table 7.2, the presence of fourth order serial correlation is unlikely to be explained by the discrepencies.

Table 7.2

Parameter Estimates of (7.1a) and (7.1b)

Parameter	Equation					
	7.1a			7.1b		
	O.L.S.	M.A.	A.R.	O.L.S.	M.A.	A.R.
b_1	454 (2.1)	289 (1.1)	-13 (0.5)	532 (2.5)	444 (1.7)	-12.4 (0.5)
b_2	-2 (0.5)	-9 (1.1)	-3 (0.9)	-5 (1.0)	-9 (1.3)	-3 (1.0)
b_3	-5 (1.0)	-7 (1.0)	-3 (0.9)	-5 (1.1)	-11 (1.1)	-3 (1.0)
b_4	0.0693 (6.0)	0.0510 (3.4)	0.0127 (0.8)	0.0732 (6.9)	0.0601 (4.1)	0.0247 (1.5)
b_5	-0.0238 (1.7)	-0.0072 (0.4)	0.0536 (1.0)	-0.0300 (2.3)	-0.0205 (1.2)	0.0313 (0.5)
b_6	-416 (2.4)	-275 (1.3)	-228 (1.7)	-475 (2.8)	-397 (1.9)	-390 (2.8)
b_7	0.2679 (2.8)	0.3610 (4.8)	0.4162 (7.7)	0.2353 (2.8)	0.3015 (4.7)	0.3478 (7.0)
α_4		0.9996 (3.7)	0.9890 (23.9)		0.9999 (4.3)	0.9999 (23.8)
σ^2	85.97	64.92	45.25	85.89	67.31	48.51

Table 7.3

Parameter Estimates for (7.7a) and (7.7b)

	(7.7a)		(7.7b)	
	O.L.S.	ARMAX	O.L.S.	ARMAX
$\hat{c}_1$	331 (13.5)	337 (8.2)	304 (4.9)	324 (6.4)
$\hat{c}_2$	49.2 (4.2)	39.7 (3.9)	46.3 (3.8)	38.9 (3.5)
$\hat{c}_3$	26.2 (3.2)	25.9 (3.8)	19.0 (1.7)	21.8 (2.1)
$\hat{c}_4$	13.5 (2.5)	12.4 (2.6)	12.5 (2.2)	11.7 (2.2)
$\hat{c}_5$	0.1659 (6.8)	0.1447 (7.4)	0.1589 (6.2)	0.1425 (6.6)
$\hat{c}_6$	0.0653 (1.9)	0.0879 (3.0)	0.0509 (1.3)	0.0768 (2.1)
$\hat{c}_7$	39.5 (4.7)	43.4 (6.4)	36.2 (4.0)	43.4 (5.4)
$\hat{c}_8$	0.5957 (10.5)	0.5931 (13.9)	0.7417 (4.4)	0.6710 (4.1)
$\hat{c}_9$			-0.1097 (0.9)	-0.0565 (0.5)
$\hat{\alpha}_2$		0.4546 (1.9)		-0.3820 (1.5)
$\hat{\alpha}_3$		-0.5414 (2.1)		-0.6173 (2.3)
$\hat{\sigma}^2$	79.0	48.0	79.5	52.8

There is not as large a variation between the O.L.S. and ARMAX estimates for (7.7a) and (7.7b) as for (7.1a) and (7.1b), but it is noticeable that $(\frac{p}{w}Y)_t$ becomes significant if autocorrelation is accounted for : as well there is a clear preference for (7.7a) over (7.7b) after ARMAX estimation. The insignificance of $\hat{\alpha}_2$ agrees with the magnitude of the t statistic for the second lag of the correlograms in Table 7.1 and again there is evidence of a unit root in the M.A.. When one recalls the frequency of this occurrence throughout the literature [107] [14] it is clear that unless some explanation may be found in the future the attractiveness of ARMAX estimators will be limited. Returning to Table 7.3 why is it that the differences are less pronounced for (7.7a) and (7.7b) than for (7.1a) and (7.1b)? This question becomes important in that there is an A.R. in the former but not the latter, and autocorrelation has generally been regarded as a serious problem only in this instance. The answer must lie in the difference between R^2 (0.999 and 0.956): for the two equation types (7.7a) and (7.7b) there is such a high signal/noise ratio that the O.L.S. estimates will be close to the M.L. estimates (see Section 4.4) and little is gained by taking account of autocorrelation (in terms of parameter estimates), but to the extent that the residual variance is a measure of predictive accuracy superior forecasts should be forthcoming from the ARMAX versions of all equations.

Patterns of autocorrelation in quarterly (and monthly) econometric models are unlikely to be of the simple first order type so beloved of many empirical researchers[7], and this establishes the major advantage

[7] In this connection it is important to note that the Durbin-Watson statistic was satisfactory for the consumption equations.

stemming from the possession of an ARMAX estimator i.e. the added flexibility gained by the ability to contemplate a wider variety of specifications - perhaps even an A.R.-M.A. in the disturbance term. Until more is known about the likely form that disturbances take once attention is shifted from yearly models, this flexibility is crucial to good estimation and prediction.

7.3.2 Stochastic Behavioural Models

Our second application of ARMAX estimators will be discussed under the heading of "stochastic behavioural relations". These derive from the assumptions embodied in (7.2) and (7.4) and are concerned with the origin of the disturbance term in (7.7a) and (7.7b). Frequently the relations (7.2) and (7.4) are estimated as equations in their own right and disturbance terms are then appended, suggesting that it is natural to include these from the beginning of the exercise and to derive the reduced form under these assumptions. If this is done the resulting error term is an amalgam of $(\frac{p_t}{w_t})$ and i.i.d. variables : a specification which cannot be dealt with in a satisfactory manner by any existing estimator. Therefore a slightly different set of conditions concerning the introduction of disturbance terms will be adopted viz. (7.3) and (7.4) will become

$$L_t^* = A(\frac{p_t}{w_t})^{\beta} Y_t + e_1(t) \qquad (7.11)$$

$$L_t = W(L)\, L_t^* + e_2(t) \qquad (7.12).$$

The equations state that the potential demand for labour may fluctuate as a result of stimuli from factors other than prices and output, and that actual labour flows adjust to potential demand in a stochastic fashion. This type of model arises in a number of

econometric applications e.g. see Williams [118] for an example from the consumer durables field, so that a solution to this type of model has applications in other areas. Following the R.B.A. W(L) is taken as a rational function and the reduced form of (7.11) and (7.12) is

$$V(L)L_t = U(L)\left[A\left(\frac{p_t}{w_t}\right)^{\beta} Y_t \right] + U(L)\, e_1(t) + V(L)\, e_2(t) \qquad (7.13).$$

Now Chapter 6 was concerned with estimation when the error term was of the composite variety in (7.13) and an estimator (PH/R) was presented to provide an efficient solution. Equations (7.7a) and (7.7b) are now re-estimated with PH/R and PH/U (under the restriction that $e^* = 0$) in order to discover if there are stochastic elements in the behavioural relations. Table 7.4 contains the statistics for a judgement (with $\lambda = \sigma_2^2/\sigma_1^2$), where the order of the polynomials V(L) and U(L) are the same as in the preceding section in order to enable a direct comparison with the results of that section. In practice the structure of V(L) and U(L) should be determined again for the new specification.

A comparison of the residual variances in Tables 7.3 and 7.4 evinces the conclusion that allowance for stochastic elements in (7.3) and (7.4) would not seem to be of great benefit, but that the implied restrictions are compatible with the data. The former conclusion is scarcely surprising in view of the low value for $\hat{\rho}(1)$ in Table 7.1, and it is certain that a more complex set of assumptions is required e.g. U(L) might be of second order with $u_o = 0$. Nevertheless it has been demonstrated that the techniques of Chapter 6 may be applied in the context of an econometric model, and that instances may arise when a derived equation has the distinctive disturbance format that PH/R was designed to accommodate.

Table 7.4

Parameter Estimates for (7.7a) and (7.7b)

Stochastic Behavioural Relations

Estimate	(7.7a)		(7.7b)	
	PH/R	PH/U	PH/R	PH/U
$\hat{c}_1$	360 (5.5)	360 (5.5)	358 (5.3)	332 (7.2)
$\hat{c}_2$	41.4 (4.3)	41.4 (4.3)	41.0 (4.1)	43.3 (4.1)
$\hat{c}_3$	28.6 (2.4)	28.6 (2.4)	27.5 (2.4)	23.0 (2.7)
$\hat{c}_4$	13.8 (3.0)	13.8 (3.0)	14.0 (2.7)	12.9 (2.5)
$\hat{c}_5$	0.1534 (6.8)	0.1534 (6.8)	0.1536 (6.6)	0.1554 (7.6)
$\hat{c}_6$	0.0943 (2.9)	0.0943 (2.9)	0.0924 (2.8)	0.0733 (2.3)
$\hat{c}_7$	42.7 (4.2)	42.7 (4.3)	41.5 (3.9)	39.2 (5.3)
$\hat{c}_8$	0.5658 (8.3)	0.5658 (8.3)	0.5924 (5.8)	0.6503 (7.7)
$\hat{c}_9$			-0.0237 (0.4)	0.0019 (0.01)
$\hat{\alpha}_1$	0.3665	0.3668 (2.0)	0.3529	-0.0510 (0.8)
$\hat{\alpha}_2$			0.0028	-0.3509 (1.7)
$\hat{\lambda}$	0.17 (0.9)		0.17 (0.9)	
$\hat{\sigma}^2$	74.17	74.17	76.29	74.88

7.3.3 Rational Lags

Rational lags give rise to ARMAX models in two ways. Firstly, as demonstrated in Chapter 1, the M.A. and A.R. transfer functions will be identical. Secondly, as outlined in the preceding section, if one wishes to allow the behavioural relations to possess a stochastic element the implementation of rational lags will result in a model that calls for the use of PH/R. Although little attention has been paid to the latter theme the former is a central feature in a number of articles e.g. [13] [91].

7.4 Final Summary and Conclusion

Having explored some of the general areas of research likely to become important in the future it is now time to summarize the findings of the thesis. As these have been presented in the conclusions to each chapter only a brief summary is contained here.

1. Of all the estimators considered the Phillips/Box Jenkins type emerged as the most useful, although it appeared that samples of around 70 would be required if the familiar properties of M.L. estimators were to become operative. However, although limited, Monte Carlo studies indicated that this conclusion should be qualified if there was non-stationarity in the exogenous variables, leading to the conclusion that more investigation of the impact of the exogenous regressors is called for.

2. Of the alternatives proposed to 1 only the frequency domain versions are likely to be worth pursuing as these estimators require quantities e.g. the periodogram, that are of interest in themselves, and offer the possibility of computational savings if the sample is large. To decide the full extent of their applicability it will be necessary to assess

their performance in the estimation of more complex models e.g. Model F of Chapter 3, than were adopted in Chapter 4. As indicated there the computational load will be heavy.

3. Extensions to systems of equations may be performed relatively easily in both domains. The next step in this direction must be toward simultaneous relations, and this will pose serious algorithmic obstacles in that the Phillips approach is not well suited to F.I.M.L. estimation. As a first step one may adapt the framework of Chapter 5 to derive an analogue to 3SLS. Perhaps this will be where the frequency domain emerges victorious as the development of a F.I.M.L. version should be relatively straightforward once all data has been Fourier-transformed.

4. Estimators may be designed within an ARMA framework to solve a signal extraction problem in which the signal follows an ARMA process. It was found that such a model arose in a number of areas in economics and statistics although the quoted examples were limited in their scope e.g. in their derivation of an optimal seasonal filter Hannan et al. [43 eq. (25) p. 38] demonstrate the need to estimate an ARMA model that is similar in structure to one generated by a components model.[8] It will be necessary to tidy the approach up if the method comes into vogue, particularly in the root finding routines.

5. As Chapter 7 attempted to show the availability of ARMAX estimators endows an investigator with a good deal of flexibility and allows a wider range of model to be entertained for the data. This has obvious

---

[8] The algorithms of Chapter 6 could well be applied as the authors sought estimates by trial and error.

importance in relation to complex autocorrelation patterns but also for incorporating various non-linearities that may arise in econometric model building. Nevertheless, in order to exploit the full potential of ARMAX estimators, there must be a wider investigation into suitable test statistics for specification decisions - especially in view of the suspicion that the correlogram may be a fairly blunt instrument in small samples. To date there are only two sets of Monte Carlo studies available [11] [74] with the former being conducted on samples of 200 - far beyond what is typically available to econometricians - while the other, performed twenty years ago in the era of desk calculators, found that the distribution of the Q statistic $\left(= N \sum_{j=1}^{k} \hat{\rho}^2(j)\right)$ departed substantially from χ^2 in samples of size 15 and 35.

In the present chapter a number of examples have been communicated in order to illustrate that time series techniques can enrich econometrics. The thesis has touched on only a small array of these methods, and most attention has been lavished on a demonstration that it is relatively easy to integrate them into a traditional econometric framework. Yet, to an appreciable extent, the number and degree of sophistication of estimators currently available exceeds the tools proposed for discriminating between alternative models, so that in the future the most fruitful research will be in the realms of specification and diagnostics rather than estimation.

BIBLIOGRAPHY

[1] Aigner, D.J., "A Compendium on Estimation of the Autoregressive-Moving Average Model from Time Series Data", International Economic Review, 12 (1971), 348-369.

[2] Aitchison, J. and S. Silvey, "Maximum Likelihood Estimation Procedures and Associated Tests of Significance", Journal of the Royal Statistical Society, 22(1960), 154-167.

[3] Akaike, H., "On a Successive Transformation of Probability Distributions and its Application to the Analysis of the Optimum Gradient Method", Annals of the Institute of Statistical Mathematics (Tokyo), 11(1959), 1-16.

[4] Anderson, T.W., The Statistical Analysis of Time Series (John Wiley, New York, 1971).

[5] Bailey, M.J., "Prediction of an Autoregressive Variable Subject both to Disturbances and to Errors of Observation", Journal of the American Statistical Association, 60(1965), 164-181.

[6] Barnard, G.A., "Statistical Inference", Journal of the Royal Statistical Society, Series B, 11(1949), 115-139.

[7] Baumol, W., Economic Dynamics (MacMillan, London, 2nd ed. 1959).

[8] Beveridge, W.H., "Weather and Harvest Cycles", Economic Journal, 31(1921), 429-452.

[9] Box, G.E.P. and G.M. Jenkins, Time Series Analysis : Forecasting and Control (Holden-Day, San Francisco, 1970).

[10] Box, G.E.P., G.M. Jenkins and D.W. Bacon, "Models for Forecasting Seasonal and Non-Seasonal Time Series", in Advanced Seminar on Spectral Analysis of Time Series, ed. by B. Harris (John Wiley, New York, 1967).

[11] Box, G.E.P. and D.A. Pierce, "Distribution of Residual Autocorrelations in Autoregressive - Integrated Moving Average Time Series Models", Journal of the American Statistical Association, 64(1970), 1509-1526.

[12] Box, G.E.P. and G.C. Tiao, "A Change in Level of a Non-Stationary Time Series", Biometrika, 52(1965), 181-192.

[13] Bray, J., "Dynamic Equations for Economic Forecasting with the G.D.P. - Unemployment Relation and the Growth of G.D.P. in the U.K. as an Example", Journal of the Royal Statistical Society, Series A, 134(1971), 167-227.

[14] Brewer,K.R.W., V.W. Fitzgerald and P.J. Hagen, "Adaptive Seasonal Adjustment - Some Numerical Results", (Paper read to the First National Conference of Australian and New Zealand Economists, Melbourne, 1970).

[15] _____ , "Some Consequences of Temporal Aggregation of Time Series Models", (Mimeo).

[16] Cagan, P., "The Monetary Dynamics of Hyper-Inflation", in Studies in the Quantity Theory of Money, ed. by M. Friedman (University of Chicago Press, Chicago, 1956).

[17] Conte, S.D., Elementary Numerical Analysis (McGraw-Hill, New York, 1965).

[18] Couts, D., D. Grether and M. Nerlove, "Forecasting Non-Stationary Economic Time Series", Management Science, 13(1966), 1-22.

[19] Cramer, H. and M.R. Leadbetter, Stationary and Related Stochastic Processes (John Wiley, New York, 1967).

[20] Crockett, J.B. and H. Chernoff, "Gradient Methods of Maximization", Pacific Journal of Mathematics, 5(1955), 33-50.

[21] Dhyrmes, P., Econometrics: Statistical Foundations and Applications (Harper and Row, New York, 1970).

[22] Dhrymes, P., L. Klein and K. Steiglitz, "Estimation of Distributed Lags", International Economic Review, 11(1970), 235-250.

[23] Durbin, J., "Efficient Estimation of Parameters in Moving-Average Models", Biometrika, 46(1959), 306-316.

[24] _____ "The Fitting of Time - Series Models", Review of the International Statistical Institute, 28(1960), 233-244.

[25] _____ "Efficient Fitting of Linear Models for Continuous Stationary Time Series from Discrete Data", Bulletin of the International Statistical Institute, 38(1961), 273-282.

[26] Fishman, G., Spectral Methods in Econometrics (Harvard University Press, Cambridge, 1969).

[27] Freudenstein, F. and B. Roth, "Numerical Solution of Systems of Nonlinear Equations", Journal of the American Association of Computing Machinery, 10(1963), 550-556.

[28] Friedman, M., A Theory of the Consumption Function (Princeton University Press, Princeton, 1957).

[29] Gantmacher, F.R., The Theory of Matrices (Chelsea, New York, 1960).

[30] Goldberger, A.S., Econometric Theory (John Wiley, New York, 1964).

[31] Goldsmith, R., A Study of Saving in the United States (Princeton University Press, Princeton, 1956).

[32] Golub, G.H., "Numerical Methods for Solving Linear Least Squares Problems", Numerische Mathematik, 7(1965), 206-216.

[33] Granger, C.W.J., "The Typical Spectral Shape of an Economic Variable", Econometrica, 34(1966), 150-161.

[34] Grenander, U. and G. Szebö, Toeplitz Forms and their Applications (University of California Press, Berkeley, 1957).

[35] Grether, D.M., Studies in the Analysis of Economic Time Series (Unpublished doctoral thesis, Stanford University, 1969).

[36] Grether, D.M. and M. Nerlove, "Some Properties of 'Optimal' Seasonal Adjustment", Econometrica, 38(1970), 682-704.

[37] Haavelmo, T., "The Statistical Implications of a System of Simultaneous Equations", Econometrica, 11(1943), 1-12.

[38] Hall, R.E., "Transform Methods for the Study of Distributed Lags", Working Paper No. 132 in Economic Theory and Econometrics, University of California, Berkeley, 1968.

[39] Hammersley, J.M. and D.C. Handscomb, Monte Carlo Methods (Methuen, London, 1964).

[40] Hannan, E.J., Multiple Time Series (John Wiley, New York, 1970).

[41] ———, "The Estimation of a Changing Seasonal Pattern", Journal of the American Statistical Association, 59(1964), 1063-1077.

[42] ———, "Measurement of a Wandering Signal Amid Noise", Journal of Applied Probability, 4(1967), 90-102.

[43] Hannan, E.J., R.D. Terrell and N.Tuckwell, "The Seasonal Adjustment of Economic Time Series", International Economic Review, 11(1970), 24-52.

[44] ———, "The Estimation of Mixed Moving Average-Autoregressive Systems", Biometrika, 56(1969), 206-224.

[45] Hannan, E.J. and D.F. Nicholls, "The Estimation of Mixed Regression, Autoregression, Moving Average and Distributed Lag Models", Cowles Foundation Discussion Paper No. 291, Yale University, 1970.

[46] ———, "The Identification Problem for Multiple Equation Systems with Moving Average Errors", Econometrica (forthcoming).

[47]————————, Time Series Analysis (Methuen, London, 1960).

[48] Hartley, W., "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares", Technometrics, 3(1961), 269-280.

[49] Hendry, D.F. and P.K. Trivedi, "Maximum Likelihood Estimation of Difference Equations with Moving Average Errors : A Simulation Study", London School of Economics and Political Science Discussion Paper No. 7, 1970.

[50] Henshaw, R.C., "Application of the General Linear Model to Seasonal Adjustment of Economic Time Series", Econometrica, 34(1966), 381-395.

[51] Holmes, J., "A Condition for Independence of Permanent and Transitory Components of a Series", Journal of the American Statistical Association, 66(1971), 13-16.

[52] Holt, C.C., "Forecasting Trends and Seasonals by Exponentially Weighted Moving Averages", O.N.R. Memorandum No. 52, Carnegie Institute of Technology, 1957.

[53] Howrey, P., "Dynamic Properties of a Condensed Version of the Wharton Model", (Paper presented at the Conference on Research in Income and Wealth, National Bureau of Economic Research, 1969).

[54] I.B.M. System / 360 Scientific Subroutine Package (360A-CM-03x), Version III, Programmers' Manual.

[55] Jenkins, G.M. and D.G. Watts, Spectral Analysis and its Applications (Holden-Day, San Francisco, 1968).

[56] Johnston, J., Econometric Methods (McGraw-Hill, New York, 1963).

[57] Jorgenson, D.W., "Rational Distributed Lag Functions", Econometrica, 34(1966), 135-149.

[58] ————————, "Anticipations and Investment Behaviour", in The Brookings Quarterly Econometric Model of the United States ed. by J. Duesenberry, G. Fromm, L.R. Klein and E. Kuh (North-Holland, Amsterdam, 1965).

[59] Jorgenson, D.W., J. Hunter and M.I. Nadiri, "A Comparison of Alternative Econometric Models of Quarterly Investment Behaviour", Econometrica, 38(1970), 187-213.

[60] Kendall, M.G., Contributions to the Study of Oscillatory Time Series (Cambridge University Press, Cambridge, 1946).

[61] Kendall, M.G. and A. Stuart, The Advanced Theory of Statistics Vol. III (Griffin, London, 1966).

[62] Kmenta, J. and R.F. Gilbert, "Estimation of Seemingly Unrelated Regressions with Autoregressive Disturbances", Journal of the American Statistical Association, 65(1970), 186-197.

[63] Koyck, L.M., Distributed Lags and Investment Analysis (North-Holland, Amsterdam, 1954).

[64] Lawton, W.H. and E.A. Sylvestre, "Elimination of Linear Parameters in Nonlinear Regression", Technometrics, 13(1971), 461-467.

[65] Leenders, G.T., "Adaptive Forecasting Under Simple Probabilistic Assumptions", Report No. 6406 of the Econometric Institute, Netherlands School of Economics, 1964.

[66] Leuthold, R.M., A.J.A. MacCormick, A. Schmitz and D.G. Watts, "Forecasting Daily Hog Prices and Quantities: A Study of Alternative Forecasting Techniques", Journal of the American Statistical Association, 65(1970), 90-107.

[67] de Leeuw, F., "The Demand for Capital Goods by Manufacturers: A Study of Quarterly Time Series", Econometrica, 30(1962) 407-423.

[68] Levenberg, K., "A Method for the Solution of Certain Nonlinear Problems in Least Squares", Quarterly of Applied Mathematics, 2(1944), 164-168.

[69] Liviatan, N., "Consistent Estimation of Distributed Lags", International Economic Review, 4(1963), 44-52.

[70] Maddala, G.S. and A.S. Rao, "Maximum Likelihood Estimation of Solow's and Jorgenson's Distributed Lag Models", Review of Economics and Statistics, 53(1971), 80-88.

[71] Malinvaud, E., Statistical Methods of Econometrics (North Holland, Amsterdam, 2nd ed. 1970).

[72] Mann, H.B. and A. Wald, "On the Statistical Treatment of Linear Stochastic Difference Equations", Econometrica, 11(1943), 173-220.

[73] Marquardt, D.W., "An Algorithm for Least Squares Estimation of Non-Linear Parameters", S.I.A.M. Journal, 11(1963), 431-441.

[74] Matthai, A. and M.B. Kannan, "The Applicability of Large Sample Tests for Moving Average and Autoregressive Schemes to Series of Short Length - an Experimental Study; Partl: Moving Averages", Sankhya, 11(1951), 218-238.

[75] Mood, A.M. and F.A. Graybill, Introduction to the Theory of Statistics (McGraw-Hill, New York, 2nd ed. 1963).

[76] Nadiri, M.I. and S. Rosen, "Interrelated Factor Demand Functions", American Economic Review, 59(1969), 457-471.

[77] Naylor, T.N., J.L. Balintfy, D.S. Burdick and K. Chu, Computer Simulation Techniques (John Wiley, London, 1968).

[78] Nelson, C., "Joint Estimation of Parameters of Correlated Time Series", Report No. 7052 of the Center for Mathematical Studies in Business and Economics, University of Chicago, 1971.

[79] ————, "The Prediction Performance of the FRB-MIT-Penn. Model of the U.S. Economy", Report No. 7117 of the Center for Mathematical Studies in Business and Economics, University of Chicago, 1971.

[80] Nerlove, M., The Dynamics of Supply (Johns Hopkins, Baltimore, 1958).

[81] ————, "Distributed Lags and Unobserved Components", in Ten Economic Studies in the Tradition of Irving Fisher (John Wiley, New York, 1967), 127-169. ed. W. Feller.

[82] ————, "On the Structure of Serial Dependence in Some U.S. Price Series", Report No. 7102 of the Center for Mathematical Studies in Business and Economics, University of Chicago, 1971.

[83] Nerlove, M. and S. Wage, "On the Optimality of Adaptive Forecasting", Management Science, 10(1964), 207-224.

[84] Nerlove, M, and K.J. Arrow, "A Note on Expectations and Stability", Econometrica, 26(1958), 297-305.

[85] Nicholls, D., Inference Problems for Vector Linear Time Series Models (Unpublished Doctoral Dissertation, Australian National University, 1971).

[86] Norton, W.E. and J.R. Broadbent, Equations for Personal Consumption Expenditure (Reserve Bank of Australia, Occasional Paper No. 3B, 1970).

[87] Norton, W.E., K.E. Schott and K.M. Sweeny, Employment and Prices (Reserve Bank of Australia, Occasional Paper No. 3F, 1971).

[88] Pagan, A.R., "The Estimation of Models with Moving Average Disturbance Terms", (Paper read to the First National Conference of Australian and New Zealand Economists, Melbourne, 1970).

[89] Parzen, E., "Time Series Analysis for Models of Signals plus White Noise", in Advanced Seminar on Spectral Analysis of Time Series, ed. by B. Harris (John Wiley, New York, 1967).

[90] Phillips, A.W., "The Estimation of Systems of Difference Equations with Moving Average Disturbances", (Paper read to the Econometric Society Meeting, San Francisco, 1966).

[91] Pierce, D.A., "Least Squares Estimation in the Regression Model with Autoregressive - Moving Average Error Terms", Biometrika, 58(1971), 299-312.

[92] Powell, M.J.D., "A Survey of Numerical Methods for Unconstrained Optimization", S.I.A.M. Review, 12(1970), 79-97.

[93] Quenouille, M., "A Large Sample Test of the Goodness of Fit of Autoregressive Schemes", Journal of the Royal Statistical Society, Series A., 110(1947), 123-129.

[94] ————————, "Discrete Autoregressive Schemes with Varying Time Intervals", Metrika, 1(1958), 21-27.

[95] Ralston, A., A First Course in Numerical Methods (McGraw-Hill, New York, 1965).

[96] Sargan, J.D., "Wages and Prices in the United Kingdom : A Study in Econometric Methodology", in Econometric Analysis for National Economic Planning, ed. by P.E. Hart, G. Mills and J.K. Whitaker (Butterworths, London, 1964).

[97] Scadding, J.L., "Adaptive Expectations in the Context of Optimal Estimation Techniques", Technical Report No. 20, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1969.

[98] Schmitz, A. and D.G. Watts, "Forecasting Wheat Yields: An Application of Parametric Time Series Modelling", American Journal of Agricultural Economics, 52(1970), 247-254.

[99] Shanno, D.F., "An Accelerated Gradient Projection Method for Linearly Constrained Nonlinear Estimation", S.I.A.M. Journal of Applied Mathematics, 18(1970), 322-334.

[100] Steiglitz, K. and L. McBride, "A Technique for the Identification of Linear Systems", I.E.E.E. Transactions on Automatic Control, 10(1965), 461-464.

[101] Steiglitz, K. and R.E. Rogers, "Maximum Likelihood Estimation of Rational Transfer Function Parameters", I.E.E.E. Transactions on Automatic Control, 12(1967), 594-597.

[102] Struik, A., "On the Viability of Box-Jenkins Seasonal Models", (Paper read to the Second National Conference of Australian and New Zealand Economists, Sydney, 1971).

[103] Telser, L., "Discrete Samples and Moving Sums in Stationary Stochastic Processes", Journal of the American Statistical Association, 57(1962), 348-368.

[104] Theil, H. and S. Wage, "Some Observations on Adaptive Forecasting", Management Science, 10(1964), 198-206.

[105] Thomas, J.J. and K.F. Wallis, "Seasonal Variation in Regression Analysis", Journal of the Royal Statistical Society, Series A, 134(1971), 57-72.

[106] Tintner, G., Econometrics (John Wiley, New York, 1952).

[107] Trivedi, P.K., "Inventory Behaviour in U.K. Manufacturing, 1956-67", Review of Economic Studies, 37(1970), 517-527.

[108] Turnovsky, S.J., "Some Empirical Evidence on the Formation of Price Expectations", Discussion Paper No. 143 Department of Economics, Wharton School of Finance and Commerce, University of Pennsylvania, 1969.

[109] Wagle, B., "Some Techniques of Short-Term Sales Forecasting", The Statistician, 16(1966), 253-273.

[110] Walker, A.M., "The Asymptotic Distribution of Serial Correlation Coefficients for Autoregressive Processes with Dependent Residuals", Proceedings of the Cambridge Philosophical Society, 50(1954), 60-64.

[111] ————, "Some Consequences of Superimposed Error in Time Series Analysis", Biometrika, 47(1960), 33-43.

[112] ————, "Large - Sample Estimation of Parameters for Moving - Average Models", Biometrika, 48(1961), 343-357.

[113] ————, "Large-Sample Estimation of Parameters for Auto-regressive Processes with Moving-Average Residuals", Biometrika, 49(1962), 117-131.

[114] ————, "Asymptotic Properties of Least Squares Estimates of Parameters of the Spectrum of a Stationary Non-Deterministic Time Series", Journal of the Australian Mathematics Society, 4(1964), 363-384.

[115] Whittle, P., "Estimation and Information in Stationary Time Series", Arkiv för Mathematik, 2(1954), 423-434.

[116] ————, "Gaussian Estimation in Stationary Time Series", Bulletin of the International Statistical Institute, 39(1962), 105-129.

[117] ————————, Prediction and Regulation, (The English Universities Press, London, 1963).

[118] Williams, R.A., "Demand for Consumer Durables: Stock Adjustment Models and Alternative Specifications of Stock Depletion", Review of Economic Studies (forthcoming).

[119] Wold, H., A Study in the Analysis of Stationary Time Series (Almquist and Wicksell, Uppsala, 2nd ed. 1954).

[120] Wright, C., "Estimating Permanent Income: a Note", Journal of Political Economy, 73(1969), 845-850.

[121] Yaglom, A.M., An Introduction to the Theory of Stationary Random Functions, (Prentice-Hall, Englewood Cliffs, New Jersey, 1962).

[122] Yule, G.U., "On a Method for Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspots Numbers", Philosophical Transactions of the Royal Society of London, Series A, 226(1927), 267-298.

[123] Zellner, A., "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias", Journal of the American Statistical Association, 57(1962), 348-368.

[124] ————————, "Econometric Estimation with Temporally Dependent Disturbance Terms", International Economic Review, 2(1961), 164-178.

[125] ————————, "On the Analysis of First Order Autoregressive Models with Missing Data", International Economic Review 7(1966), 72-76.

[126] Zellner, A. and M.S. Geisel, "Analysis of Distributed Lag Models with Applications to Consumption Function Estimation", Econometrica, 38(1970), 865-888.

[127] Zellner, A. and C.M. Montmarquette, "A Study of Some Aspects of Temporal Aggregation Problems in Econometric Analysis", Review of Economics and Statistics, 53(1971), 335-342.