ALTERNATIVE APPROACHES TO TREND ESTIMATION

A THESIS

SUBMITTED IN PARTIAL FULFILMENT

OF THE REQUIREMENTS OF

SHEFFIELD HALLAM UNIVERSITY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

by Stephen James Salter

MAY 1996

ABSTRACT

This thesis suggests a general approach for estimating the trend of a univariate time series. It begins by suggesting and defining a set of "desirable" trend properties, namely "Fidelity", "Smoothness", "Invariance" and "Additivity", which are then incorporated into the design of an appropriate non-stationary time series model.

The unknown parameters of the model are then estimated using a wide selection of "optimal" procedures, each parameter having at least two such procedures applied to it. Attention is paid to the development of algorithms to implement the procedures in practice.

The model is gradually extended from a basic, non-seasonal model consisting of a simple lagged trend to a general, seasonal model incorporating a variable parameter, general autoregressive trend.

ACKNOWLEDGEMENTS

My thanks go to my supervisors, Professor Warren Gilchrist and Doctor Bekia Fosam, whose positive suggestions and criticisms have led to many improvements, and who have patiently and diligently kept me focused throughout the thesis, especially over the last year, ensuring my sights remained fixed on the main task of writing up the material.

CONTENTS

INTRODUCTION

The main aim of this introductory chapter is to put the reader in the correct frame of mind for what follows in the later chapters.

The ideas which motivated the thesis did not begin, and will doubtless not end, with the thesis itself, but have evolved over, what is now, a period of about fifteen years. In this respect the thesis is simply a suitable vehicle, which came along at the right time, and which provided the means by which I might illustrate the application of what are a set of more general and fundamental concepts.

Hence, in order to fully understand what this thesis attempts to achieve and equally why I felt such an attempt was important, it is necessary to appreciate the events which led up to it. In doing so I need to address, (albeit from a slightly different angle), some fairly basic "time series" concepts and although I can sympathise with the more specialist reader for having to go over what will be familiar territory, I make no apologies, since I feel it is essential that the all readers are aware of "where I'm coming from" to use contemporary vernacular.

0.1 THE ORIGINS

The origins of this thesis lie in the late seventies, a period during which I worked for both Industry and the Government as a practising statistician. From my experience at least, the aspect of time series analysis which most concerned both sectors was the evaluation and forecast of trends, although each sector adopted a quite different approach to the problem.

Their different approaches, which coloured their different concepts of trend, were essentially practical and were decided by which particular computer package had been chosen for their main-frame computer. (It

should be remembered that at this time the Personal Computer, or at least a P.C. of sufficient power, was not commercially available).

The Government stood firmly by the traditional form of time series decomposition in using the X-11 seasonal adjustment program developed by its American counterparts, (Shishkin et al, 1967), at the Bureau of the Census, (1969), (see also Den Butter and Fase, 1991), whilst Industry was increasingly being converted from the earlier exponentially-based methods of Brown, (1959, 1962) and Winters, (1960), to the more sophisticated ARIMA adopted by the Box-Jenkins package, (1976), (see Bowerman and O'Connell, (1979), for an excellent dual account of these).

Also for completeness I should perhaps mention two other approaches, namely Spectral Analysis, (see Priestley, 1981, for a comprehensive exposition), which had some success in specialised areas, and State Space techniques, whose basis lay in the Kalman filter, (Kalman, 1960, 1963), and which was under development in applications such as Bayesian Forecasting, (Harrison and Stevens, 1976), (see Abraham and Ledolter, 1983, and Hamilton, 1994, for excellent treatments).

As previously mentioned, my main concern was the inconsistent, and what I felt was inadequate, way in which both the X-11 and Box-Jenkins models evaluated the trend, which it is instructive to review.

0.11 THE X-11 APPROACH

This disaggregates the each time series observation, y_t, into its constituents of trend, x_t, seasonality, s_t, and residual, e_t, as follows,

$$y_t = x_t + s_t + e_t \qquad (0.01)$$

The trend values and seasonalities x_t, are essentially calculated by repeatedly smoothing y_t, using moving averages, to give x_t, seasonally smoothing $y_t - x_t$ to give s_t and then adjusting the s_t to produce no

overall aggregated yearly seasonality. (Minor variations on this theme are adopted for the multiplicative model, $y_t = x_t.s_t.e_t$).

The point is that there is no particular model, as such, for the trend, which is purely defined as a moving average which immediately raises the problems of end-effects and forecasting, (points which are addressed to a certain extent by Dagum, 1975, 1980).

These inadequacies are confirmed by the Government's reluctance to publish the trend figures. Instead it provides, what are known as, seasonally adjusted values, which are the values of $y_t - s_t$. Note from equation (0.01) that these are also the values of $x_t + e_t$, i.e. the values of the trend plus its residual. Given the large residuals in many Government series, these values can be very misleading to the uninitiated.

0.12 THE BOX-JENKINS APPROACH

Here, the observed series, z_t, is expressed in terms of "k" of its previous values, z_{t-1}, z_{t-1}, ... , z_{t-k}, and an independent residual, e_t, i.e.

$$z_t = \vartheta_1.z_{t-1} + \vartheta_2.z_{t-1} + \ldots + \vartheta_d.z_{t-d} + e_t \qquad (0.02)$$

In terms of the backward operator $\mathbb{B}$, such that $\mathbb{B}^k.z_t = z_{t-k}$, this can be written,

$$\Theta_k(\mathbb{B}).z_t = (1-\vartheta_1.\mathbb{B} - \vartheta_2.\mathbb{B}^2 - \ldots - \vartheta_k.\mathbb{B}^k).z_t = e_t \qquad (0.03)$$

The model also contains the constraint that the series z_t is stationary which in terms of equation (0.03) means that z_t can be written as,

$$z_t = \Theta_k(\mathbb{B})^{-1}.e_t \qquad (0.04)$$

In other words that the function $\Theta_k(\mathbb{B})$ can be inverted. If it cannot Box-Jenkins suggest differencing the original series y_t, (i.e. successively applying the difference operator $(1-\mathbb{B})$ to y_t), until this is the case. Hence after differencing "d" times, the model becomes,

$$(1-\mathbb{B})^d.y_t = z_t = \Theta_k(\mathbb{B})^{-1}.e_t \qquad (0.05)$$

Note that there is no specific mention of either the trend, x_t, or, in the more general model, the seasonality, s_t. In the case of equation (0.05) it is recovered as the difference between the non-stationary series, y_t, and the stationary series, z_t, i.e.

$$x_t = y_t - z_t = (1-(1-\mathbb{B})^d).y_t \qquad (0.06)$$

This differencing operation innately assumes that the trend is polynomial in nature, which is constricting, but also has some rather odd side-effects, when we try and relate it to the model of equation (0.01).

Suppose, for example that the true model of the observations can be written as a simple polynomial in time plus an independent residual, i.e.

$$y_t = b_0 + b_1.t + b_2.t^2 + e_t \qquad (0.07)$$

Differencing this model three times to achieve stationarity gives,

$$(1-\mathbb{B})^3.y_t = z_t = (1-\mathbb{B})^3.e_t \qquad (0.08)$$

However the function $(1-\mathbb{B})^3$ is not invertible, and hence the simple polynomial model of equation (0.07) cannot be modelled using the Box-Jenkins approach.

0.13 WHAT IS TREND?

From the last two sections we see that an X-11 user defines trend in terms of moving averages, whereas a Box-Jenkins user would define it as the difference between a non-stationary and stationary series. Alternatively, a spectral analyst regards it as the combination of all those cycles whose time periods are longer than the series itself. In other words, whilst exact definitions exist, they are different and all peculiar to whichever form of analysis they are produced by.

As soon as we try and obtain a non-specific definition, this exactness is replaced by vagueness. For example, O'Muircheartaigh and Francis, in their statistical dictionary, (O'Muircheartaigh and Francis, 1981), begin to describe trend as "*The broad underlying movement of a time series*".

This raised the question of whether one could draw up, if not an exact general definition, then at least a set of desirable properties which a trend should possess. If this could be done, then the general estimation of trend could be approached from its definition and/or properties and not arrived at as a bi-product of other techniques.

0.14 DESIRABLE TREND PROPERTIES

With the above in mind I drew up an initial set of the following four properties.

0.141 Fidelity

This property attempts to put the words "*broad underlying movement*" into some more exact form. It basically says that the trend values, x_t, and the original observations, y_t, should not deviate too much from each other for any significant period of time.

On a very simple level this could, for example, suggest that a function of the form,

$$\text{Fidelity} = \sum (y_t - x_t)^2 \qquad (0.09)$$

is kept as small as possible.

0.142 Smoothness

This property states that the trend series itself should follow a reasonably smooth curve, which has no sudden jumps. To me this seems intuitively obvious, since if I am presented with an apparent trend which is not "smooth", I find myself still having to mentally smooth the series to obtain its *underlying movement*.

Nevertheless it must be said that there are many statisticians who would dispute this criterion of smoothness; largely because the time series analysis technique they happen to use, produces a *trend* as one of its bi-products, which could not, under any liberal interpretation, be considered to be smooth.

However, I do have, or rather did have, one great proponent, the late Sir Maurice Kendall, who said "*The essential idea of trend is that it shall be smooth, which in practice means that we should like to represent it by a continuous and differentiable function of the time*", (Kendall, 1973).

Again, on a very simple level this could, for example, suggest that a function of the form,

$$\text{Smoothness} = \sum (x_t - x_{t-1})^2 \qquad (0.10)$$

is kept as small as possible in the case of a discrete time series.

In fact we can go further and generalise the definition from first differences to differences of order "d" by writing it as,

$$\text{Smoothness} = \sum (\nabla^d . x_t)^2 \qquad (0.11)$$

$$\text{where } \nabla . x_t = (1 - \mathbb{B}) . x_t = x_t - x_{t-1}$$

0.143 Invariance

If the original observed time series, y_t, consisted of values which exactly fell on a straight line, i.e. $y_t = b_0 + b_1.t$, it would seem fairly obvious that the resulting trend values, x_t, should also fall on the same straight line, i.e. that $x_t = b_0 + b_1.t$ as well.

Stated formally, this would mean that the model should be invariant to straight line data. Extending this property would imply that the trend model should be invariant to as many simple functions as possible, notably the polynomial family, (since other "well-behaved" functions can be approximated by them, using Taylor's theorem).

In fact, if the Fidelity and Smoothness example suggestions of sections **0.141** and **0.142** are adopted, we can see that this property is automatically taken account of to a certain extent.

0.144 Additivity

This states that if the general trend value of a particular observed series of "T" values, $y_1, y_2, \ldots, y_T$, is defined as the function $x_t(y_1, y_2, \ldots, y_T)$ and for another observed series of "T" values, $z_1, z_2, \ldots, z_T$, as $w_t(z_1, z_2, \ldots, z_T)$, then the trend function $xw_t(y_1+z_t, y_2+z_t, \ldots, y_T+z_T)$ based on the observed series, $y_1+z_1, y_2+z_2, \ldots, y_T+z_T$, should be such that $xw_t = x_t + w_t$.

Note that if the form of the trend function is linear, then,

$$x_t = \alpha_0 + \alpha_1.y_1 + \alpha_2.y_2 + \ldots + \alpha_T.y_T \qquad (0.12)$$

$$w_t = \beta_0 + \beta_1.z_1 + \beta_2.z_2 + \ldots + \beta_T.z_T \qquad (0.13)$$

$$xw_t = \gamma_0 + \gamma_1.(y_1+z_1) + \gamma_2.(y_2+z_2) + \ldots + \gamma_T.(y_t+z_T) \qquad (0.14)$$

and the additivity property is satisfied if the γ coefficients satisfy,

$$\gamma_0 = \alpha_0+\beta_0 \text{ and } \gamma_k = (\alpha_k.y_k+\beta_k.z_k)/(y_k+z_k) \text{ for } 1 \leq k \leq T \qquad (0.15)$$

In other words this condition enables us to test whether particular trend models are consistent for aggregated series,(such as total costs), and their disaggregations, (such as labour and material costs).

0.15 A BASIC TREND MODEL

Weighting together equations (0.09) and (0.10) we obtain the following function, φ, defined as:

$$\varphi = (1-p).\sum_{t=1}^{T}(x_t - y_t)^2 + p.\sum_{t=2}^{T}(x_t - x_{t-1})^2 \qquad (0.16)$$

where $0 \leq p \leq 1$

Hence by minimising φ, it would appear that we would satisfy properties **0.141**, **0.142** and, to a lesser extent **0.143**. This was not a particularly new function as I had experimented with it previously whilst developing other smoothing methods. By varying the smoothing parameter, p, a time series could be smoothed to any degree one fancied. However, at the time, minimising φ appeared to be just another smoothing algorithm.

In fact, I was not the only one who had experimented with φ, since, as it later turned out, Whittaker, (1923, 1924), had been doing the same thing almost sixty years earlier.

However what Whittaker missed, or at least did not follow up, was what, with dramatic simplicity, changes equation (0.16) from yet another smoothing algorithm to a complete time series model, thus opening it up fully to areas such as estimation, (with which this thesis is mainly concerned), inference and forecasting.

All this requires is that the differences $y_t - x_t$ and $x_t - x_{t-1}$ be defined as random variables e_t and a_t, thus,

$$y_t = x_t + e_t \tag{0.17}$$

$$x_t = x_{t-1} + a_t \tag{0.18}$$

Equations (0.17) and (0.18), which might be termed the "Fidelity" and "Smoothness" equations, constitute, what is termed in this thesis as, the "Basic" model, which, we shall be extending as the thesis develops. For example, we will be generalising the "Smoothness" equation of (0.18) to have an autoregressive structure given by,

$$x_t = \vartheta_1 . x_{t-1} + \vartheta_2 . x_{t-2} + \ldots + \vartheta_d . x_{t-d} + a_t \tag{0.19}$$

in which the autoregressive parameters, ϑ_i, are either pre-specified, (fixed parameter model), usually to satisfy invariance properties, or need to be estimated, (variable parameter model).

0.16 ESTIMATION

As we shall see, estimation of the trend values, x_t, will, for the "Basic" model of equations (0.17) and (0.18), also require the estimation of the variances of e_t, (σ_e^2), and a_t (σ_a^2), which we shall refer to as "residual" variances. In addition, for the "General", (variable parameter), model, which utilises (0.19) in place of (0.18), we shall also require estimates of the autoregressive parameters, i.e. (ϑ_i), the set of ϑ_i, for i=1 to d.

The philosophy behind the process of estimation in this thesis is to investigate as many different procedures as possible. However, in doing so, we have deliberately limited our search to estimators that are produced as a result of some form of "optimal" process, (on the assumption that this will, in consequence, result in estimators which possess "optimal" properties).

THE APPROACH TO TREND ESTIMATION

STAGE ONE: DEFINE THE TREND PROPERTIES

(I) FIDELITY **(II) SMOOTHNESS**

(III) INVARIANCE **(IV) ADDITIVITY**

STAGE TWO: INCORPORATE THE PROPERTIES INTO A MODEL

$$y_t = x_t + e_t$$

$$x_t = \vartheta_1 . x_{t-1} + \vartheta_2 . x_{t-2} + \ldots + \vartheta_d . x_{t-d} + a_t$$

STAGE THREE: ESTIMATE THE MODEL PARAMETERS

(A) THE TREND VALUES (x_t)

(I) USING MINIMISATION OF WHITTAKER'S FUNCTION {CHAPTER 1}

(II) USING GENERALISED LEAST SQUARES REGRESSION {CHAPTER 1}

(III) USING THE STATE SPACE APPROACH {CHAPTERS 2,3}

(B) THE RESIDUAL VARIANCES (σ_e^2, σ_a^2)

(I) USING MAXIMUM LIKELIHOOD {CHAPTERS 4,5}

(II) USING MINIMUM VARIANCE OF QUADRATIC FORMS {CHAPTERS 4,5}

(C) THE AUTOREGRESSIVE PARAMETERS (ϑ_i)

(I) USING LEAST SQUARES {CHAPTER 7}

(II) USING MAXIMUM LIKELIHOOD {CHAPTER 7}

The table on the previous page summarises the estimation procedures which we shall be using within the context of the overall trend estimation approach. In the next section we go into a little more detail regarding the structure of the thesis.

0.2 THESIS OUTLINE

0.21 CHAPTER ONE

The chapter begins with a matrix formulation and minimisation of Whittaker's function of equation (0.16) to produce trend, (x_t), estimates, going on to look at some results for different values of the smoothness parameter "p".

We then move on to the "Basic" model described by equations (0.17) and (0.18), and, after consideration of distributional assumptions, estimate the trend values using Classical, Generalised Least Squares, (GLS), regression.

We continue by showing how the above two processes will produce identical trend estimates if the smoothness parameter is suitably interpreted in terms of the model's residual variances.

Finally we conclude by discussing the limitations of the basic model with respect to the "Invariance" property and hence introduce the, more general, autoregressive trend of (0.19), which overcomes these.

0.22 CHAPTER TWO

Chapter two is exclusively concerned with the State Space approach to solving the general autoregressive model introduced in the previous chapter.

After formulating the model in State Space format, and introducing the main ideas of the method, it goes on to apply the usual three-stage procedure of prediction, filtering and smoothing, albeit without the

need for Normality assumptions, to produce another set of trend estimates.

Being essentially a Bayesian technique, no investigation of State Space methodology would be complete without consideration of prior distributions, and the chapter ends by looking at what a "vague prior" would imply for the State Space starting values; producing a result required in the next chapter.

0.23 CHAPTER THREE

The chapter again concerns itself with State Space methodology, and begins by showing that, given the assumption of a vague prior, the State Space estimates and those of the Classical approach of chapter one produce identical results.

The rest of the chapter examines the stages of the State Space procedure in more detail, ending with a little more insight into its interpretation of smoothness.

0.24 CHAPTER FOUR

Chapter four concerns itself wholly with the theory behind the estimation of the variances of the residuals, e_t, and a_t, which necessitates the re-introduction of a Normality assumption.

It begins by showing how Classical and State Space approaches both lead to the same Likelihood function, which is then maximised to produce a set of equations, whose solution gives the residual variance estimates.

It then goes on to consider an alternative, Minimum Variance approach which utilises some known results relating to the variance of the quadratic form, showing that this gives rise to same equation set as those produced by Maximum Likelihood.

The chapter finishes with a short section highlighting the difference

between the optimal variance estimates produced earlier in the chapter and those "intuitively" based on mean squared residuals.

0.25 CHAPTER FIVE

Chapter five continues the work on residual variance estimation by showing how the optimal variance equations of chapter four could be efficiently solved in practice, by taking advantage of the different ways by which they were produced. It concludes with a general algorithm for their estimation.

0.26 CHAPTER SIX

In chapter six the non-seasonal model, so far discussed, is extended to include the effects of seasonality, adapting the techniques of the previous chapters to reproduce their main results for the seasonal case.

0.27 CHAPTER SEVEN

Chapter seven is concerned with the estimation of autoregressive parameters, (i.e. the variable parameter model), for both the non-seasonal and seasonal cases. Two estimation procedures are dealt with, (Least Squares and Maximum Likelihood), and the conditions under which their estimates are identical is addressed.

0.28 CHAPTER EIGHT

Chapter eight is a short chapter containing some suggestions for further work, discussing limiting forms, model extensions, inference, forecasting, data trials and multivariate models.

0.29 CONCLUSION

The thesis concludes with a short summary of its achievements and limitations.

0.3 THE TREND MODEL IN HISTORICAL CONTEXT

Whilst there are to my knowledge no examples of applying the direct approach, suggested by this thesis, to the problem of trend estimation, we are able, by tracing the development of Whittaker's function in (0.16), to see how models similar to the ones adopted by this thesis, have been applied in other areas.

As has already been mentioned earlier in the chapter, I was not the first person to have considered the merits of Whittaker's function and so I did not expect that I would have been the first to investigate the model of equations (0.17) and (0.18). I was not disappointed in this, and in this section I take the reader through those events which were either related to or utilised the formats of either equation (0.16) or equations (0.17) and (0.18).

0.31 WHITTAKER'S SMOOTHING FUNCTION

The story begins on the 14th November 1919, at the Edinburgh Mathematical Society, where Professor E. T. Whittaker read a paper entitled 'On a New Method of Graduation'. It took four years for the paper to be published, (Whittaker, 1923), which suggests it underwent substantial revision. A year later Whittaker included the idea in a small part of a chapter entitled 'Graduation, or the Smoothing of Data' in a treatise (Whittaker and Robinson, 1924) on numerical mathematics.

One of the interesting points regarding both the paper and the book are the pains to which the authors went to show how a solution, once derived, could be implemented. Page upon page is devoted to performing a series of example calculations, which transform data from one column into another and then into another and so on in a series of simple stages in order to arrive at the required result.

As we shall see this updating process is fundamental to the philosophy of State Space modelling. Whittaker used it to solve, albeit approximately, (he actually needed to choose a relatively high value

for "p" of about 0.95 in equation (0.16) to achieve his result), what was in essence a minimisation problem and whilst we certainly cannot attribute the ideas of updating (or recursive filtering or feedback control, to give it two of its modern synonyms) to Whittaker, the fact that he was able to effect any solution at all to his *problem of graduation* is because the methods of updating and minimisation are, in reality, just two sides of the same coin.

The equivalence of updating and minimisation, or State Space and Classical approaches to the problem of graduation or smoothing one of the topics of this thesis. Familiar bells begin to ring, when it is appreciated that State Space modelling is essentially a Bayesian technique since, as Lindley demonstrates in his books on Probability and Inference, (Lindley, 1965a, 1965b), most Classical techniques can be shown to be special cases of their Bayesian counterparts, (see also Lindley, 1972).

When Whittaker explained his *method of graduation*, he never actually used the word *trend* as such, although he did say *"....there is a strong antecedent probability that if the observations had been more accurate the curve would have been smooth ."* In other words he addressed his ideas to the scenario of an underlying smooth process whose observations exhibited erratic behaviour because of measurement errors. I shall be taking up this point later. It is essentially what is meant by the a State Space approach, and, although it is highly unlikely that he fully appreciated the potential of what he had hinted at, his words were nevertheless interestingly prophetic.

However, Whittaker did something much more concrete than unwittingly hint at what was to become a standard time series technique. He suggested a compromise. The two most important words he used in his paper were *smoothness* and *fidelity*. His *trend* was to be a compromise between its smoothness and its fidelity to the original data. On the one hand he could represent the trend by a horizontal line drawn through the middle of the series of observations. This would certainly be smooth but would in no way reflect any inherent movement in its level. At the other extreme he could faithfully reproduce every little

twist and turn of the original data and regard the trend as synonymous with the original series. There would be no concession to smoothness in this case. The weighting would be one hundred percent on fidelity, but obviously, in choosing this trivial extreme, the trend would retain one hundred percent of the series' observed erratic behaviour.

His compromise was to have something in between, by weighting together the smoothness and the fidelity, which in turn meant that he had to be able to measure both of them. His measurement of fidelity was straightforward. He required some function of the differences between the observed values and the resulting trend. He referred to these differences as *measurement errors*. Not surprisingly he chose squared differences for probably much the same reasons that the variance function is chosen as a measure of variability i.e. it is the simplest even function which utilises all the data and is also differentiable.

A smoothness function was not quite so straightforward. In Whittaker's words *"We may make the somewhat vague word 'smooth' more precise by interpreting it to mean, e.g. that the third differences of the derived series are to be very small."* Why third differences? The obvious choice would have been the simplest case of first differences. As we shall see later the choice of first differences produces some very odd results with certain types of series since it is only invariant for a set of observations with a constant mean, (see section **0.143**).

Whittaker knew what he was doing. He'd probably tried out the first differences case on some simple series and realised its limitations, and in doing so further realised, that to be able to successfully avoid this type of problem arising in a practical situation, he would need to move up to a third level of differencing.

The reasons for Whittaker's choice will become evident as this thesis develops. For Whittaker to have been aware of this possible pitfall meant that he either, had tremendous insight, or that he, or his assistant, spent many long nights with slide rule or log tables. To be realistic, the truth probably lay somewhere in between.

In order to aid understanding in the initial stages of chapter one of this thesis, the simpler definition of smoothness using first differences is adopted and hence we will refer to equation (0.16) as Whittaker's equation and define φ in this equation to be Whittaker's smoothing function. The model of equations (0.17) and (0.18) is referred to as the basic model. Later on in the chapter, the definition is extended to a more general function and hence a more general model, which will include both first differences and Whittaker's third differences as special cases.

0.32 WAHBA'S SPLINE SOLUTION

Through the seventies, the theory of spline functions was successively developed by Wahba and others, (in order of development Reinsch, 1967, Kimmeldorf and Wahba, 1970, 1971, Wahba and Wold, 1975, Wahba, 1977, Golub, Heath and Wahba, 1979, and Craven and Wahba, 1979), and applied, using a generalised method of cross-validation, to Whittaker's problem, giving values, not only for each x_t, but for the value of "p" in equation (0.16) and the value of "d" in equation (0.11) also. The drawback was that the method was computationally $O(T^3)$, i.e. the number of calculations required increased, (at least), in proportion to the cube of the number of data points.

0.33 SHILLER'S SMOOTHNESS PRIORS

In 1973 a paper by Shiller, (Shiller, 1973), investigated the prediction of a known time series y_t from another known series z_t using distributed lags, i.e. a relationship of the form:

$$y_t = \beta_0.z_t + \beta_1.z_{t-1} + \ldots + \beta_L.z_{t-L} + \varepsilon_t \qquad (0.20)$$

The problem with the usual regression solution to the problem was that it produced β coefficient values that had a seemingly random pattern, which made no sense from an econometric point of view.

His solution was to place a constraint on the β_i, such that their

differences, $\nabla^n.\beta_i$, had a prior Normal distribution. This, coupled with an assumption of Normality for the ε_t, gave a posterior distribution, and hence a likelihood function, for the β_i. Given certain variance parameters, he could then maximise the likelihood to find estimates for the β_i.

Thus, the technique of choosing a prior distribution for model parameters in order to ensure their resulting smoothness was initiated and the term "smoothness priors" was born. (One can can regard smoothness priors as a special case of Bayesian priors, initially introduced, (Good, 1965), and developed, (Good and Gaskins, 1980), by Good).

0.34 THE WORK OF AKAIKE

Akaike built on the work of Shiller, developing his ideas, (Akaike, 1979a, 1979b, 1979c, 1980a, Akaike and Ishiguro, 1980), within the framework of a model incorporating trend and seasonality, i.e. $y_t = x_t + s_t + \varepsilon_t$. By placing smoothness priors, (in difference form), on the trend, x_t, and seasonal, s_t, parameters, he produced a likelihood function which could be numerically searched to give, not only parameter values but also their asymptotic variances, (often referred to as hyper-parameters, see Akaike, 1980c).

He was able to go even further since he had, in his armoury, his, (AIC), information criterion, (Akaike, 1973, 1974, Kitagawa and Akaike, 1978), which he used to determine the "optimal" order of non-seasonal and seasonal differencing, (Akaike, 1980b). Akaike's approach reduced the computational complexity from $O[T^3]$ to $O[T^2]$.

0.35 THE STATE SPACE APPROACH OF GERSCH AND KITAGAWA

During the seventies an approach to modelling time series had been developing, (Weinert, 1979), which was based on the filtering methods of Kalman, (1960, 1963). It was termed State Space modelling since it purported to encompass the whole of a time series' previous history within (the Space of) a small number of current, (State), parameters.

Kitagawa, who worked for Akaike in Tokyo, realised that Kalman's equations were efficient, well-behaved and above all adaptable and subsequently, (Kitagawa, 1981), formulated Akaike's ideas into State Space form, a form which was computationally O[T].

Independently, Gersch, (Brotherton and Gersch, 1981), reached the same conclusion, three thousand miles away in Hawaii, and also made the connection with Whittaker's work, something that Shiller, Akaike and Kitagawa had missed.

What followed, either by chance or design, were the first, (Kitagawa and Gersch, 1982), (Gersch and Kitagawa, 1983b), (Kitagawa and Gersch, 1984), of several joint papers, by Gersch and Kitagawa, (while both were A.S.A. fellows at the U.S. Bureau of Census), which developed and refined the smoothness priors approach, by incorporating additional refinements such as trading-day adjustments etc.

0.36 FURTHER DEVELOPMENTS

The latter half of the 1980's saw the utilisation of the smoothness priors approach to several models, whose application had, previously, been limited to stationary series, e.g. non-linear regression, (Shiller, 1984, Eubank, 1986), transfer functions, (Gersch and Kitagawa, 1984, 1989), multivariate models, (Gersch and Kitagawa, 1983a, Gersch, 1989, Gersch and Stone, 1990), distributed lags, (Thurman et al, 1986, Polasek, 1990), and non-stationary covariance structures using spectral estimation, (Kitagawa and Gersch, 1985a, 1985b).

0.37 ADDITIONAL MATERIAL

In addition to the specific topics covered above, papers (Gersch, 1987, Gersch and Kitagawa, 1988, Kohn and Ansley, 1988, Terasvirta et al, 1988) contain informative general expositions.

CLASSICAL APPROACHES TO TREND ESTIMATION

1.1 A MATRIX SOLUTION TO WHITTAKER'S PROBLEM

In this chapter we begin by investigating the solution of, what will be referred to as, Whittaker's problem, discussed in the last chapter. Let us restate the problem identified by equation (0.16) in the introduction:

Given a sequence of T observations $y_1, y_2, \ldots, y_t, \ldots, y_T$ we need to find the corresponding estimates, $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_t, \ldots, \hat{x}_T$ of a sequence of trend values $x_1, x_2, \ldots, x_t, \ldots, x_T$ which minimise the function φ where:

$$\varphi = (1-p).\sum_{t=1}^{T} (x_t - y_t)^2 + p.\sum_{t=2}^{T} (x_t - x_{t-1})^2 \qquad (1.01)$$

$$\text{and} \quad 0 \leq p \leq 1 \qquad (1.02)$$

This is equivalent to minimising ψ where:

$$\psi = \sum_{t=1}^{T} (x_t - y_t)^2 + (1/\omega).\sum_{t=2}^{T} (x_t - x_{t-1})^2 \qquad (1.03)$$

$$\text{where } \omega = (1-p)/p, \text{ and hence } 0 \leq \omega \leq \infty \qquad (1.04)$$

We can regard ω as a sort of odds ratio. As the proportional smoothness, p, increases from zero to one, the value of ω drops from infinity to zero. Thus the higher the value of ω, the nearer Whittaker's estimates, $\hat{x}_t$, will be to the original series y_t.

1.11 WHITTAKER'S PROBLEM IN MATRIX FORM

We shall find it useful to define the 'fidelity' or the measurement

errors, e_t, given by the difference between the observed series, y_t, and the required trend, x_t, as:

$$e_t = y_t - x_t \qquad (t=1,2,\ldots,T) \qquad (1.05)$$

and the 'smoothness' or structural errors, a_t, given by the differences between successive trend values, x_t and x_{t-1}, i.e.

$$a_t = x_t - x_{t-1} \qquad (t=2,3,\ldots,T) \qquad (1.06)$$

Hence we could write (1.03) as:

$$\psi = \sum_{t=1}^{T} e_t^2 + (1/\omega).\sum_{t=2}^{T} a_t^2 \qquad (1.07)$$

Writing out the set of T equations (1.05) in full we have,

$$\begin{aligned} e_T &= y_T - x_T \\ &\ldots\ldots\ldots\ldots \\ e_2 &= y_2 - x_2 \\ e_1 &= y_1 - x_1 \end{aligned}$$

or, more conveniently, in vector terms, as:

$$\underline{e} = \underline{y} - \underline{x} \qquad (1.08)$$

where

$$\begin{aligned} \underline{y}^T &= (y_T, \ \ldots\ , y_2, y_1) \\ \underline{x}^T &= (x_T, \ \ldots\ , x_2, x_1) \\ \underline{e}^T &= (e_T, \ \ldots\ , e_2, e_1) \end{aligned}$$

Similarly writing out the set of T-1 equations (1.06) in full we have,

$$
\begin{aligned}
a_T &= x_T - x_{T-1} \\
&\cdots\cdots\cdots\cdots \\
a_3 &= x_3 - x_2 \\
a_2 &= x_2 - x_1
\end{aligned}
$$

and again in vector terms, as:

$$\underline{a} = \mathbf{D}.\underline{x} \tag{1.09}$$

where: $\underline{a}^T = (a_T, \ \ldots\ , a_3, a_2)$

and **D** is the T-1 × T difference matrix with structure:

$$\mathbf{D} = \begin{bmatrix} +1, -1, 0, 0, 0, 0, \ldots\ldots\ldots \\ 0, +1, -1, 0, 0, 0, \ldots\ldots\ldots \\ 0, 0, +1, -1, 0, 0, \ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{etc} \end{bmatrix}$$

Hence (1.03) or (1.07) can be written,

$$\psi = (\underline{y} - \underline{x})^T.(\underline{y} - \underline{x}) + (1/\omega).\underline{x}^T.\mathbf{D}^T.\mathbf{D}.\underline{x} \tag{1.10}$$

1.12 OPTIMISATION

Because ψ in (1.10) or more clearly (1.07) is a weighted sum of squared errors, the value of $\hat{\underline{x}}$ obtained by minimising ψ can be thought of as a *least squares* estimate or perhaps, more correctly, a *weighted least squares* estimate.

Differentiating ψ in (1.10) with respect to $\underline{x}$, and setting the result equal to zero gives:

$$(2/\omega).\mathbf{D}^T.\mathbf{D}.\hat{\underline{x}} = 2.(\underline{y} - \hat{\underline{x}})$$

i.e. $$(1/\omega).\mathbf{D}^T.\mathbf{D}.\hat{\underline{x}} = (\underline{y} - \hat{\underline{x}}) \tag{1.11}$$

where $\hat{\underline{x}}$ is the T×1 vector of Whittaker's estimates, i.e.

$$\hat{\underline{x}}^T = \{\hat{x}_T, \ \ldots \ , \hat{x}_2, \hat{x}_1\} \qquad (1.12)$$

Rearranging (1.11), we get:

$$\underline{y} = (I_T + D^T.D/\omega).\hat{\underline{x}} \qquad (1.13)$$

where I_T is the T × T identity matrix.

Hence Whittaker's estimate, $\hat{\underline{x}}$, of $\underline{x}$, is given by:

$$\boxed{\hat{\underline{x}} = (I_T + D^T.D/\omega)^{-1}.\underline{y}} \qquad (1.14)$$

a new, efficient solution of which follows:

1.13 AN EFFICIENT SOLUTION TO WHITTAKER'S PROBLEM

Whittaker's estimate, $\hat{\underline{x}}$, given by (1.14), can be rearranged to give:

$$(\omega.I_T + D^T.D).\hat{\underline{x}} = \omega.\underline{y} \qquad (1.15)$$

Writing the equations out in full gives:

$$\begin{aligned}
(1+\omega).\hat{x}_1 - \hat{x}_2 &= \omega.y_1 \quad (1)\\
-\hat{x}_1 + (2+\omega).\hat{x}_2 - \hat{x}_3 &= \omega.y_2 \quad (2)\\
- \hat{x}_2 + (2+\omega).\hat{x}_3 - \hat{x}_4 &= \omega.y_3 \quad (3)\\
-\hat{x}_{t-1} + (2+\omega).\hat{x}_t - \hat{x}_{t+1} &= \omega.y_t \quad (t)\\
- \hat{x}_{T-1} + (1+\omega).\hat{x}_T &= \omega.y_T \quad (T)
\end{aligned}$$

....(1.16)

The problem now is, to use these T equations, to find the T unknowns $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T$. One obvious solution is to use Gaussian Elimination to solve the set of linear equations. Teodorescu uses this to solve the same problem, (*Teodorescu,1989*). However, there is a better way of proceeding which not only solves the problem more efficiently, but also gives us an interesting insight into what happens when we come to view the situation through State Space glasses.

The first thing to notice is that, apart from the first equation and the last, all the equations are of the form:

$$-\hat{x}_{t-1} + (2+\omega).\hat{x}_t - \hat{x}_{t+1} = \omega.y_t \qquad 1 < t < T \tag{1.17}$$

The characteristic equation of the recursive relationship on the left hand side of this equation i.e. between the $\hat{x}$'s is:

$$\lambda^2 - (2+\omega).\lambda + 1 = 0 \tag{1.18}$$

This is an important equation which will be met in a different context in chapter two. It has solutions:

$$\lambda = 1+\omega/2 \pm \sqrt{(1+\omega/2)^2-1} \tag{1.19}$$

Since $\omega/2 < (1+\omega/2)^2-1 < (1+\omega/2)^2$, for $\omega > 0$, then the smallest of the roots, call it λ^*, where:

$$\boxed{\lambda^* = 1+\omega/2 - \sqrt{(1+\omega/2)^2-1}} \tag{1.20}$$

must lie between 0 and 1. The largest root, equal to $1/\lambda^*$, must, since their product is unity, therefore be greater than one. Note that the following relationships exist between ω and λ^*.

$$2+\omega = \lambda^* + 1/\lambda^* \qquad \text{and} \qquad \omega = (1-\lambda^*)^2/\lambda^* \tag{1.21}$$

We can therefore write equation (1.17) as:

$$-\hat{x}_{t-1} + (\lambda^* + 1/\lambda^*).\hat{x}_t - \hat{x}_{t+1} = \omega.y_t = ((1-\lambda^*)^2/\lambda^*).y_t \tag{1.22}$$

for $1 < t < T$, where λ^* is defined as in (1.20).

The next important step is to introduce a new variable $\hat{f}_t$, which in the light of what is to follow in chapter two, may be termed a pseudo-filtered value. This is defined, for $1 < t < T$, as:

$$\hat{f}_t = (\hat{x}_t - \lambda^*.\hat{x}_{t+1})/(1-\lambda^*) = \hat{x}_t/(1-\lambda^*) - \lambda^*.\hat{x}_{t+1}/(1-\lambda^*) \tag{1.23}$$

Hence (1.22) can be written:

$$\hat{f}_t = \lambda^*.\hat{f}_{t-1} + (1-\lambda^*).y_t \tag{1.24}$$

Increasing the time period by one and rewriting (1.24) we have, now for $0 < t < T-1$,

$$\boxed{\hat{f}_{t+1} = \lambda^*.\hat{f}_t + (1-\lambda^*).y_{t+1}} \tag{1.25}$$

Again, we shall derive a very similar relationship to this in chapter two. By using the first equation of the initial set of equations (1.16), namely,

$$(1+\omega).\hat{x}_1 - \hat{x}_2 = \omega.y_1 \tag{1.26}$$

together with with (1.23) and (1.24) above, it is relatively easy to show that:

$$\hat{f}_0 = \hat{\hat{x}}_1 \tag{1.27}$$

which would imply a value for $\hat{\hat{x}}_0$ of,

$$\hat{\hat{x}}_0 = \hat{f}_0 = \hat{\hat{x}}_1 \tag{1.28}$$

Similarly, by using the last equation of the initial set (1.16), we can show that:

$$\hat{\hat{x}}_T = \hat{f}_T = \hat{\hat{x}}_{T+1} \tag{1.29}$$

The values of $\hat{\hat{x}}_0$ and $\hat{\hat{x}}_{T+1}$ are not strictly required in obtaining the solution. Again, however, their relevance will become apparent later.

We now have the apparatus necessary for calculating the trend estimates $\hat{\hat{x}}_1$, $\hat{\hat{x}}_2$, ... , $\hat{\hat{x}}_T$. Firstly, we choose a value for $\hat{f}_0$; anything will do, although in the light of (1.28), perhaps a proxy for $\hat{\hat{x}}_1$ i.e. y_1 is the best starting value.

We then use the recursive system of (1.24), namely:

$$\hat{f}_t = \lambda^*.\hat{f}_{t-1} + (1-\lambda^*).y_t \tag{1.30}$$

to successively generate values $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, ... , $\hat{f}_T$. Notice that since λ^* lies between 0 and 1, our starting value loses its influence very quickly i.e. it is a transient; so that if the initial error in choosing $\hat{f}_0$ is ε then from repeated application of (1.30), the error in $\hat{f}_T$ will be $(\lambda^*)^T.\varepsilon$, i.e. proportional to λ^* to the power T, a very small value.

We now arrange equation (1.23) to give:

$$\hat{x}_t = \lambda^* . \hat{x}_{t+1} + (1-\lambda^*) . \hat{f}_t \qquad (1.31)$$

To obtain a value for $\hat{x}_{T-1}$ we need a value for $\hat{x}_T$ which is given by (1.29) as $\hat{f}_T$. We can then successively generate the values $\hat{x}_{T-1}$, $\hat{x}_{T-2}$, $\hat{x}_{T-3}$, ... , $\hat{x}_1$ using (1.31).

Again note that since the coefficients λ^* and $(1-\lambda^*)$ both lie between 0 and 1, the effects of the starting value $\hat{f}_0$, are even further reduced i.e. since the error in $\hat{f}_T$, and hence $\hat{x}_T$, is $(\lambda^*)^T.\varepsilon$, then, from repeated application of (1.19), the error in $\hat{x}_1$ will be $(\lambda^*)^{2T-1}.\varepsilon$, a minute value.

From (1.28), we can now use this calculated value of $\hat{x}_1$ as a new starting value for $\hat{f}_0$, and repeat the process until it converges to any repeated value of $\hat{x}_t$.

Since the new error in $\hat{f}_0$ must be $(\lambda^*)^{2T-1}$ times the value of its initial error, and λ^* lies between 0 and 1, very few, if any, repeats are needed no matter how large the initial error. Also note that the process will always converge however short the length of the series.

Therefore the solution to the equations in (1.16) simply requires the recursive application of equations (1.30) and (1.31) above, equations not dissimilar to those encountered in exponential smoothing.

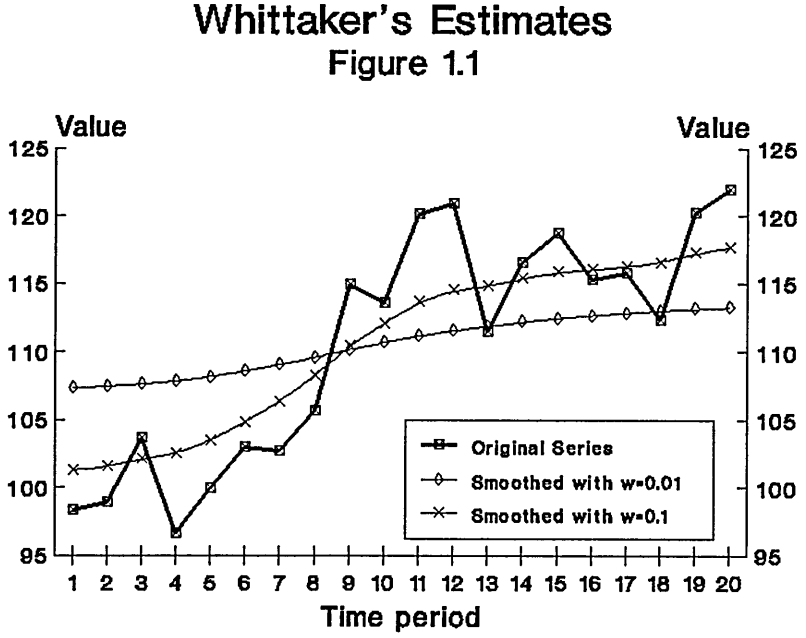
1.131 MATRIX INVERSION

Notice that the above technique also permits the general inversion of the matrix $(\mathbf{I}_T + \mathbf{D}^T\mathbf{D}/\omega)$ which we will be coming across later in equation (1.38).

If the vector of values $\mathbf{v}_i$, whose elements are all zero except for the ith which is unity, is used in place of $\underline{y}$ in (1.16), the resulting values produced for $\underline{\hat{x}}$ will be the ith column of $(\mathbf{I}_T + \mathbf{D}^T\mathbf{D}/\omega)^{-1}$.

1.14 COMPARISON OF RESULTS

Figure 1.1 shows two sets of estimates for a particular series; one with ω = 0.01, (equivalent to p=0.99), and another with ω = 0.1, (p=0.91).



We can see that as ω is increased the smoothed series moves closer to the original series and, conversely, as ω is decreased Whittaker's estimates tend to a horizontal line. This will always be true when first differences are used in the minimisation function ψ, (1.03).

As we shall see later, this leads to highly misleading estimates for certain types of series, (which Whittaker seems to have realised), and needs to be replaced with a more general form of differencing function. But more of that later.

1.2 THE INTRODUCTION OF DISTRIBUTIONAL ASSUMPTIONS

Whittaker's estimate, $\hat{\underline{x}}$, in (1.14) is only an *estimate* in a colloquial sense. To be classed as an estimate in a statistical sense, it would

need to be the realisation of an *estimator*, which is a random variable, or more precisely a vector of random variables since we are working in multivariate terms. We therefore need to introduce a stochastic element into the formulation via some distributional assumptions.

The simplest way to do this is to assume that the error terms e_t and a_t in (1.05) and (1.06) arise as realisations of random variables, E_t and A_t say, which have some simple distributional structure. The obvious choice is to have independent random variables with zero mean and constant variance, i.e. for some distribution $D_t(0,\sigma^2)$,

$$E_t \sim \text{i.d. } D_t(0;\ \sigma_e^2) \quad \underline{\text{and}} \quad A_t \sim \text{i.d. } D_{t+T}(0;\ \sigma_a^2)$$

Note that the distributions of both the E_t and A_t may change with time, whilst keeping constant respective variances. In other words it is only their means and variances which are of interest at this stage. However we do need to assume that, as well as being independent of other E_t and A_t, they are also independent of each other; in summary that.

$$\mathbb{E}[E_r.A_s] = 0 \quad \text{for all } r,s$$

$$\text{and} \qquad \mathbb{E}[E_r.E_s] = \mathbb{E}[A_r.A_s] = 0 \quad \text{if } r \neq s$$

....(1.32)

The random vector equivalent of (1.08) can be written:

$$\underline{Y} = \underline{X} + \underline{E} \tag{1.33}$$

where

$$\underline{Y}^T = (Y_T, \ \ldots\ , Y_2, Y_1)$$
$$\underline{X}^T = (X_T, \ \ldots\ , X_2, X_1)$$
$$\underline{E}^T = (E_T, \ \ldots\ , E_2, E_1)$$

Similarly, the random vector equivalent of (1.09) is:

$$\mathbf{D}.\underline{X} = \underline{A} \tag{1.34}$$

where $\underline{A}^T = (A_T, \ldots, A_3, A_2)$

Note that because of (1.34), $\underline{x}$ can not simply be regarded as a vector of unknown parameters but has to be defined as the realisations of an unknown random vector $\underline{X}$.

1.21 GENERALISED LEAST SQUARES

The realisations, of (1.33) and (1.34), can be combined to give:

$$\begin{bmatrix} \text{ø} \\ \underline{y} \end{bmatrix} = \begin{bmatrix} \mathbf{D} \\ \mathbf{I}_T \end{bmatrix} . \underline{x} + \begin{bmatrix} -\underline{a} \\ \underline{e} \end{bmatrix} \tag{1.35}$$

where ø is a conformally dimensioned vector whose elements are all zero.

Note that this ø notation is in future used for any vector or matrix whose elements are zero, and whose dimensions are usually being clear from its context.

The incorporation of the prior information of (1.34) into (1.33) is a special case of the *mixed estimation procedure* introduced by Theil and Goldberger, (*Theil and Goldberger,1961*), and further discussed in (*Theil,1970,p.346-352*).

Equation (1.35) is in the form of (A1) of appendix A, and hence we can use the results of this appendix to find the minimum mean square unbiased estimate, (MMSE), of $\underline{x}$ i.e. $\hat{\underline{x}}$. It should be noted that the residuals in vectors $\underline{a}$ and $\underline{e}$ are assumed to be uncorrelated and the covariance matrix of their joint random variate vector is given by:

$$\mathbb{C}ov\begin{bmatrix} -\underline{A} \\ \underline{E} \end{bmatrix} = \sigma_e^2 . \begin{bmatrix} \omega . I_{T-1} & \varnothing \\ \varnothing & I_T \end{bmatrix} = \begin{bmatrix} \sigma_a^2 . I_{T-1} & \varnothing \\ \varnothing & \sigma_e^2 . I_T \end{bmatrix} \qquad (1.36)$$

$$\text{where } \omega = \sigma_a^2 / \sigma_e^2$$

Hence, using (A14) of appendix A, $\hat{\underline{x}}$ is given by:

$$\hat{\underline{x}} = \left\{ \begin{bmatrix} D^T, & I_T \end{bmatrix} \begin{bmatrix} 1/\omega . I_{T-1} & \varnothing \\ \varnothing & I_T \end{bmatrix} \begin{bmatrix} D \\ I_T \end{bmatrix} \right\}^{-1} \begin{bmatrix} D^T, & I_T \end{bmatrix} \begin{bmatrix} 1/\omega . I_{T-1} & \varnothing \\ \varnothing & I_T \end{bmatrix} \begin{bmatrix} \varnothing \\ \underline{y} \end{bmatrix}$$

(Not that $(1/\omega).I_{T-1}$ is written more economically as $1/\omega.I_{T-1}$)

Therefore,

$$\boxed{\hat{\underline{x}} = \left[I_T + 1/\omega . D^T . D \right]^{-1} . \underline{y} = \Pi_T^{-1} . \underline{y}} \qquad (1.37)$$

Hence the MMSE estimate, $\hat{\underline{x}}$ in (1.37), also referred to as a Generalised Least Squares, (GLS), estimate in appendix A, is exactly the same as Whittaker's estimate in (1.14) but with ω redefined as σ_a^2 / σ_e^2, and hence, can utilise the procedure of section **1.13** for its solution.

The mean squared error of $\hat{\underline{x}}$ defined as $\mathbb{E}\left[[\hat{\underline{x}} - \underline{x}] [\hat{\underline{x}} - \underline{x}]^T \right]$ is given by (A17) of appendix A as:

$$\text{MSE}\left[\hat{\underline{x}} \right] = \sigma_e^2 . \left\{ \begin{bmatrix} D^T, & I_T \end{bmatrix} \begin{bmatrix} 1/\omega . I_{T-1} & \varnothing \\ \varnothing & I_T \end{bmatrix} \begin{bmatrix} D \\ I_T \end{bmatrix} \right\}^{-1}$$

and therefore,

$$\boxed{MSE\left[\hat{\mathbf{x}}\right] = \sigma_e^2.\left[\mathbf{I}_T + 1/\omega.\mathbf{D}^T.\mathbf{D}\right]^{-1} = \sigma_e^2.\Pi_T^{-1}} \qquad (1.38)$$

1.22 STANDARDISED RESIDUALS

We have seen that when ω is set equal to the variance ratio, σ_a^2/σ_e^2, Whittaker's estimates become MMSE estimates, i.e. GLS estimates, extended to accommodate the fact that the unknown parameters are random rather than fixed. We can see how setting ω equal to the variance ratio affects the minimisation of ψ in (1.07), which becomes:

$$\psi = \sum_{t=1}^{T} e_t^2 + \sigma_e^2/\sigma_a^2.\sum_{t=2}^{T} a_t^2$$

i.e. $$\psi = \left(\sum_{t=1}^{T} (e_t/\sigma_e)^2 + \sum_{t=2}^{T} (a_t/\sigma_a)^2 \right).\sigma_e^2$$

Eliminating the multiplier σ_e^2, which does not affect the minimisation, we get,

$$\psi^* = \psi/\sigma_e^2 = \sum_{t=1}^{T} (e_t^*)^2 + \sum_{t=2}^{T} (a_t^*)^2 \qquad (1.39)$$

where a_t^* and e_t^* are standardised residuals

Therefore minimisation of ψ^* is equivalent to the process of Ordinary Least Squares, (OLS), accompanied by its usual distributional assumptions, i.e. that the error terms a_t^* and e_t^* are independently distributed with common variance.

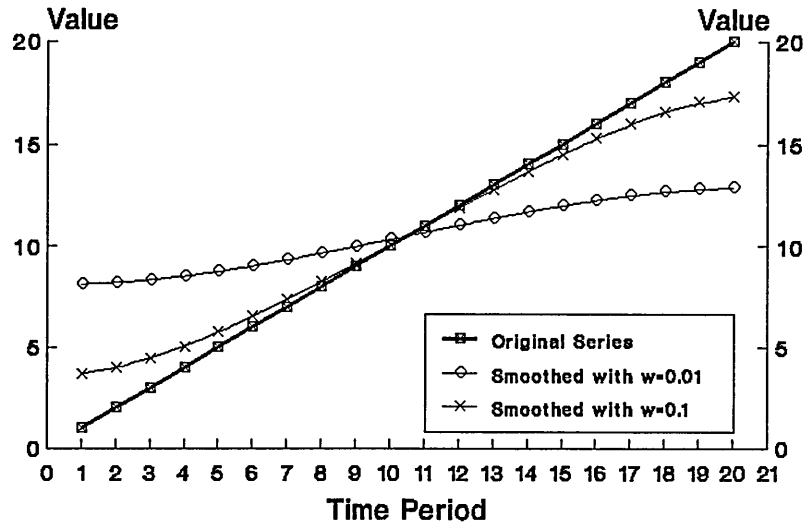
Note that (1.37) is just (A14) of appendix A, with the residuals covariance matrix given in (A17), i.e.

$$\psi^* = \begin{bmatrix} -\underline{a} , \underline{e} \end{bmatrix} \begin{bmatrix} \sigma_a^2 . I_{T-1} & \varnothing \\ \varnothing & \sigma_e^2 . I_T \end{bmatrix}^{-1} \begin{bmatrix} -\underline{a} \\ \underline{e} \end{bmatrix} \qquad (1.40)$$

1.3 LIMITATIONS OF THE MODEL: INVARIANCE

A major limitation of, what might be described as, the "Basic" model, so far presented, is highlighted by its performance on linear data, i.e. when $y_t = a + b.t$. Common sense would suggest that trend values, x_t, should also be linear in such a situation; however figure 1.2 shows this is not at all what happens. As the smoothness is increased from w=0.1, (p=0.91), to w=0.01, (p=0.99), the trend line tends to bend away from the straight line of the actual data.

The Basic Model with Straight Line Data
Figure 1.2



We can see why this should be so by reminding ourselves of the original problem of equations (1.05) to (1.07), which was to minimise ψ, where:

$$\psi = \sum_{t=1}^{T} e_t^2 + 1/\omega \sum_{t=2}^{T} a_t^2 \qquad (1.41)$$

$$\text{with } a_t = x_t - x_{t-1}, \text{ and } e_t = y_t - x_t$$

As we increase the smoothness, i.e. w is decreased, less emphasis is placed on reducing the measurement residuals e_t, and consequently more placed on reducing the structural residuals a_t, which means making x_t closer to x_{t-1} - in other words, setting x_t to a constant level.

In this respect the term "smoothness" could be more exact. What is more precisely meant is fidelity to the underlying smooth structural form, which, in this case, when a_t is defined in terms of first differences, $x_t - x_{t-1}$, is a constant level.

Thus, to be able to reproduce straight line data, (i.e. be linearly invariant), we would require a set of structural equations based on second differences, i.e. of the form, $x_t = 2.x_{t-1} - x_{t-2} + a_t$.

We can now see why Whittaker felt the need to utilise third differences in his minimisation, since this meant his model would be invariant to constant, linear and parabolic data forms, which he obviously felt were enough to describe most of the data series he had encountered.

1.31 THE GENERAL AUTOREGRESSIVE MODEL

There is, of course, no need to stop at third differences, or indeed to even limit ourselves to differences at all. We can define a general autoregressive model as:

$$x_t = \vartheta_1.x_{t-1} + \vartheta_2.x_{t-2} + \ldots + \vartheta_d.x_{t-d} + a_t \qquad (1.42)$$

which replaces (1.06), equation (1.05) remaining the same, i.e.

$$y_t = x_t + e_t \qquad (1.43)$$

The beauty of this formulation is that all the results so far produced for the "Basic" Model are equally valid for this "General" Model, since, the matrix formulations are the same for both models. The only difference is that the matrix **D**, defined as part of (1.09), is redefined as a T-d × T matrix with structure:

$$\mathbf{D} = \begin{bmatrix} 1 & -\vartheta_1 & -\vartheta_2 & \cdots & -\vartheta_d & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -\vartheta_1 & -\vartheta_2 & \cdots & -\vartheta_d & 0 & 0 & 0 & \cdots \\ 0 & 0 & \multicolumn{8}{l}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{etc}} \end{bmatrix} \qquad (1.44)$$

thus retaining the previous matrix model formulation equations of (1.08) and (1.09), but now with $\underline{a} = (a_T, a_{T-1}, \ldots a_{d+1})^T$, as:

$$\underline{e} = \underline{y} - \underline{x} \qquad (1.45)$$

$$\underline{a} = \mathbf{D}.\underline{x} \qquad (1.46)$$

1.32 INTERPRETATION OF THE STRUCTURAL EQUATION

Equation (1.42) may be written as two equations, namely,

$$\tilde{x}_t = \vartheta_1.x_{t-1} + \vartheta_2.x_{t-2} + \ldots + \vartheta_d.x_{t-d} \qquad (1.47)$$

$$x_t = \tilde{x}_t + a_t \qquad (1.48)$$

In this sense $\tilde{x}_t$ can be regarded as a one step ahead forecast of x_t, predicted from previous values, with a_t as the prediction error.

Alternatively, (1.42) could be thought of as the combination of several simpler equations, namely,

$$\begin{aligned} x_t &= \eta_1.x_{t-1} + u1_t \\ u1_t &= \eta_2.u1_{t-1} + u2_t \\ u2_t &= \eta_3.u2_{t-1} + u3_t \end{aligned}$$

$$u3_t = \eta_4 . u3_{t-1} + u4_t$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$up\text{-}1_t = \eta_p . up\text{-}1_{t-1} + up_t$$

....(1.49)

which is exactly equivalent to (1.42) if the final residual, $up_t = a_t$, and the coefficients, η_1, η_2, ... , η_p are the "p" roots of the polynomial,

$$\eta^p + \vartheta_1 . \eta^{p-1} + \ldots + \vartheta_{p-1} . \eta + \vartheta_p = 0 \qquad (1.50)$$

Note that this does not imply that the roots, and consequently the residuals, (except the last), of (1.47) have to be real. Nor does it imply that these residuals have to be stationary.

The form of equation (1.50) is addressed later, both in section **3.5** of chapter three and section **8.112** of chapter eight.

1.4 SUMMARY OF MAIN POINTS

In this chapter we have looked at, what might be termed Classical approaches of estimating the trend values, (see table on page 10 of the introduction).

In section **1.1** we addressed the minimisation of Whittaker's function and which gave us one set of "optimal" trend estimates.

In section **1.2** we introduced distributional assumptions which extended his "algorithmic" approach to that of a "Basic" model and from it produced Generalised Least Squares, (GLS), or MMSE estimates, which were shown to be identical to Whittaker's if his "ω" weighting factor was interpreted as a residual variance ratio. Section **1.3** extended the "Basic" model to give a "General" autoregressive model.

THE STATE SPACE APPROACH TO TREND ESTIMATION I

2.1 THE GENERAL AUTOREGRESSIVE MODEL

In the last chapter a Classical/Least Squares approach was applied to the estimation of the trend parameters x_t for both a "Basic" and a "General" autoregressive trend model. In this chapter we again investigate these models but using a Bayesian/State Space approach to estimate the x_t. We begin by casting the equations of the general autoregressive model, i.e. (1.42) and (1.43), in a suitable form.

The structural or smoothness equation (1.42) is written:

$$\mathbf{x}_t = \Theta.\mathbf{x}_{t-1} + a_t.\mathbf{v} \tag{2.01}$$

or in long form as,

$$\begin{pmatrix} x_t \\ x_{t-1} \\ \cdot \\ \cdot \\ x_{t-d+1} \end{pmatrix} = \begin{pmatrix} \vartheta_1, & \vartheta_2, & \cdot & \cdot & \cdot & \vartheta_d \\ 1, & 0, & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & 1, & 0, & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0, & 0, & \cdot & , & 1, & 0 \end{pmatrix} . \begin{pmatrix} x_{t-1} \\ x_{t-2} \\ \cdot \\ \cdot \\ x_{t-d} \end{pmatrix} + a_t . \begin{pmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ 0 \end{pmatrix} \tag{2.02}$$

and the measurement equation (1.43), (with the same vector $\mathbf{v}$ as above), as

$$y_t = \mathbf{v}^T.\mathbf{x}_t + e_t \tag{2.03}$$

Again, the model consists of two sets of equations, a set of structural equations, describing the process by which the trend values are generated, and a set of measurement equations, by which the values are, albeit inaccurately, observed.

The conditions placed on the structural and measurement residuals are the same as in equation (1.32) of the last chapter, i.e. that they are independently generated from distributions with zero means and constant respective variances. Thus,

$$E_t \sim \text{i.d. } D_t(0;\ \sigma_e^2) \quad \underline{\text{and}} \quad A_t \sim \text{i.d. } D_{t+T}(0;\ \sigma_a^2)$$

$$\mathbb{E}[E_r.A_s] = 0 \quad \text{for all } r,s$$

$$\text{and} \qquad \mathbb{E}[E_r.E_s] = \mathbb{E}[A_r.A_s] = 0 \quad \text{if } r \neq s$$

....(2.04)

Again because of (2.01), the trend parameters x_t must be realisations of random variables X_t.

2.2 CONDITIONAL DISTRIBUTIONS AND THEIR VARIATES

The rest of this chapter is almost exclusively concerned with conditional distributions. Because of this it is useful, before beginning, to review exactly what will be implied by this and also what notation will be used. If some of what follows seems to be unnecessarily pedantic at first, the reader is nevertheless asked to be tolerant since, unlike most situations, the results and their derivations depend crucially on what conditions are placed on particular distributions.

By conditional we mean conditional on the amount of knowledge available in defining the distribution. In this sense all parametric distributions are conditional. For example using the form B(r/n,p) for the Binomial distribution explains that the distribution of the variate "r" is conditional on the values given to the parameters "n" and "p". Thus the distribution B(r/5,0.5) is very different from B(r/100,0.003).

A more formal way of looking at this is to define the variate itself as being conditional, i.e. to regard the conditional variate r/n,p as having a Binomial distribution, or if the meaning was clear we could use a mix of the two and define the variate r/n as having a Binomial, B(p), distribution. It is this latter mixed approach that will be adopted in this chapter. Note that is far from unusual since whilst $X \sim N(\mu,\sigma^2)$ is taken to mean that the variate X has a Normal distribution with given parameters μ and σ^2, X_t(or X/t) $\sim N(\mu_t,\sigma_t^2)$ is taken to imply a Normal variate X given parameters μ_t, σ_t^2 and time parameter t.

This last description of X_t is similar to the one used in this chapter with one important addition. As well as being conditional on which distributional parameters such as μ and σ^2 and which time parameter t is assumed to be known, the distribution of the variate is also conditional on the amount of relevant data that is assumed to be known.

The relevant data are the observed values of the time series in equation (2.03), i.e. the observations y_1, y_2 etc.

The mean and variance of the distribution of any of the variates in equations (2.01) and (2.03) will vary depending on how many observations of y can be assumed to be known. Thus the distributions of X_t/y_1, $X_t/y_1,y_2$ etc. will all be different distributions.

What is more, for the conditional variate $X_t/y_1,y_2,\ldots,y_n$, "t" can be greater than, equal to or less than n, when it is referred to as a predicted, (t>n), filtered, (t=n), or smoothed, (t<n), variate. Note also that the variates Y_t, E_t and A_t, whose corresponding realisations are used in equations (2.01) and (2.03) could be referred to likewise, although we will not need to do so.

Finally a note on subscripts. T is usually reserved for the last observed y value, i.e. the time series has been observed over T time periods, n and k denote the latest value of y that can be assumed in parts of the argument, (hence $1 \leq n,k \leq T$), and t is the time period of

the variate in question. Also a rather unwieldy definition of a conditional variate such as $X_t/y_1,y_2,\ldots,y_n$ is replaced by the more economical $X_t(n)$.

2.3 THE ESSENCE OF STATE SPACE ESTIMATION

The State Space approach regards the model, not as two sets of equations per se, but as defining two sets of distributions, a set of prior distributions, (albeit indirectly), and a set of conditional distributions.

Equations (2.01) define how the prior distribution of each trend parameter, X_t, changes over time and equations (2.03) define the conditional distributions of each measured variate, Y_t, given x_t.

The successive application of, what is fundamentally Bayes rule, leads to the conditional or "posterior" distribution of each X_t given all the values of y_t, i.e. the smoothed variate $X_t(T)$. The estimate of x_t is then chosen as the mean of its posterior distribution.

In the last chapter the calculation of what turns out to be the means of $X_t(T)$ for t=1,2,...T was performed directly. In the State Space approach, however, we perform the operation in three stages, namely prediction, filtering and smoothing.

2.4 PREDICTION

In the prediction phase we relate the parameters of the filtered vector variate, $\mathbf{x}_{t-1}(t-1)$, to those of the predicted vector variate, $\mathbf{x}_t(t-1)$ using (2.01).

Suppose the filtered vector variate $\mathbf{x}_{t-1}(t-1)$, has an unspecified distribution with vector mean and covariance matrix given by:

$$\mathbf{x}_{t-1}(t-1) \sim UD(\mu_{t-1}(t-1);\ \Sigma_{t-1}(t-1)) \qquad (2.05)$$

And the predicted vector variate $\mathbf{x}_t(t-1)$, has an unspecified distribution with vector mean and covariance matrix given by:

$$\mathbf{x}_t(t-1) \sim UD(\mu_t(t-1);\ \Sigma_t(t-1)) \tag{2.06}$$

Using (2.01), the predicted values vector, $\mathbf{x}_t(t-1)$, can be written,

$$\mathbf{x}_t(t-1) = \Theta.\mathbf{x}_{t-1}(t-1) + a_t(t-1).\mathbf{v} \tag{2.07}$$

Taking expectations of (2.07), and noting that the predicted error residual, $a_t(t-1)$, must have zero mean, we have, for the means in (2.05) and (2.06),

$$\boxed{\mu_t(t-1) = \Theta.\mu_{t-1}(t-1)} \tag{2.08}$$

Also, taking covariances of both sides of (2.07), and noting that $\mathbf{x}_{t-1}(t-1)$ and $a_t(t-1)$ are independent, we get, for the covariances in (2.05) and (2.06), and the variance of $a_t(t-1)$, in (2.04),

$$\boxed{\Sigma_t(t-1) = \Theta.\Sigma_{t-1}(t-1).\Theta^T + \sigma_a^{\,2}.\mathbf{v}.\mathbf{v}^T} \tag{2.09}$$

Equations (2.08) and (2.09) are known as the prediction equations.

In a similar way to the above (2.03) can be written,

$$y_t(t-1) = \mathbf{v}^T.\mathbf{x}_t(t-1) + e_t(t-1) \tag{2.10}$$

and again by taking expectations and variances of (2.10), noting the

independence of $\mathbf{x}_t(t-1)$ and $e_t(t-1)$, and also that $e_t(t-1)$ has mean zero and variance σ_e^2, we have:

$$\mathbb{E}[y_t(t-1)] = \mathbf{v}^T.\mu_t(t-1) \tag{2.11}$$

$$\mathbb{V}ar[y_t(t-1)] = \mathbf{v}^T.\Sigma_t(t-1).\mathbf{v} + \sigma_e^2 \tag{2.12}$$

Before proceeding any further with the model we shall first need to demonstrate a result regarding conditional distributions.

2.4 THE "BEST" LINEARLY CONDITIONAL MULTIVARIATE DISTRIBUTION

Consider a vector comprising the partitioned random vectors **X** and **Y**, which has an unspecified distribution, (denoted by UD), whose mean and covariance matrix is conformally partitioned as:

$$\begin{Bmatrix} \mathbf{X} \\ \mathbf{Y} \end{Bmatrix} \sim \text{UD} \left\{ \begin{matrix} \mu_x \\ \mu_y \end{matrix} ; \begin{matrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{matrix} \right\} \tag{2.13}$$

It perhaps goes without saying that the accuracy by which we can measure the variates contained in **X**, (which currently has a <u>prior</u> distribution UD($\mathbf{x}$; Σ_{xx})), can be improved by utilising known values of **Y** as long as **X** and **Y** are correlated, i.e. $\Sigma_{xy} \neq \emptyset$. The question is how to combine these values to "best" advantage.

If the distribution is specified as Normal the situation is well-documented and, *(Drhymes,1970,p16)* or *(Harvey:1989,p.165)*, for example, show that the conditional distribution of **X**, given the vector of realisations of **Y** i.e. **y**, is given by **X(y)** where:

$$\mathbf{X}(\mathbf{y}) \sim N \left\{ \mu_x + \Sigma_{xy}.\Sigma_{yy}^{-1}.(\mathbf{y} - \mu_y) \; ; \; \Sigma_{xx} - \Sigma_{xy}.\Sigma_{yy}^{-1}.\Sigma_{yx} \right\} \tag{2.14}$$

In the general case of an unspecified distribution, we find that we can obtain the same result by looking at the problem in a different way.

Suppose we linearly regress each of the variates in the vector **X**, i.e. x_i on the values in **Y**, i.e. **y**,

$$x_i = a_i + \mathbf{b}_i^T.\mathbf{y} + e_i \tag{2.15}$$

The results of performing each of the i=1 to m regressions individually can be summarised as:

$$\mathbf{X} = \mathbf{a} + \mathbf{B}.\mathbf{Y} + \mathbf{e} \tag{2.16}$$

where a_i is the ith element of **a** and $\mathbf{b}_i^T$ is the ith row of **B**.

For the whole population of values in **X** and **Y**, the values of **a** and **B** which minimise each of the e_i in **e** are given by:

$$\mathbf{a} = \mu_x - \Sigma_{xy}.\Sigma_{yy}^{-1}.\mu_y \tag{2.17}$$

$$\mathbf{B} = \Sigma_{xy}.\Sigma_{yy}^{-1} \tag{2.18}$$

Hence combining (2.16), (2.17) and (2.18) we get:

$$\mathbf{X} = \mu_x + \Sigma_{xy}.\Sigma_{yy}^{-1}.(\mathbf{Y} - \mu_y) + \mathbf{e} \tag{2.19}$$

Note that (2.19) satisfies all the relationships in (2.13) since on taking expectations of (2.19) we get, since $\mathbb{E}[\mathbf{e}] = \emptyset$,

$$\mathbb{E}[\mathbf{X}] = \mu_x + \Sigma_{xy}.\Sigma_{yy}^{-1}.(\mu_y - \mu_y) + \emptyset = \mu_x \tag{2.20}$$

and also using (2.19),

$$\mathbb{E}[(\mathbf{X} - \mu_x).(\mathbf{Y} - \mu_y)^T] = \Sigma_{xy}.\Sigma_{yy}^{-1}.\mathbb{E}(\mathbf{Y} - \mu_y).(\mathbf{Y} - \mu_y)^T] + \mathbb{E}[\mathbf{e}.(\mathbf{Y} - \mu_y)^T]$$

which since all elements of **e** are independent of all elements of Y, simplifies to:

$$\Sigma_{xy} = \Sigma_{xy}.\Sigma_{yy}^{-1}.\Sigma_{yy} + \varnothing = \Sigma_{xy} \tag{2.21}$$

Taking covariances of (2.19) gives,

$$\Sigma_{xx} = \Sigma_{xy}.\Sigma_{yy}^{-1}.\Sigma_{yy}.\Sigma_{yy}^{-1}.\Sigma_{yx} + \Sigma_{ee} = \Sigma_{xy}.\Sigma_{yy}^{-1}.\Sigma_{yx} + \Sigma_{ee} \tag{2.22}$$

Hence the covariance matrix of the errors **e** is given by:

$$\Sigma_{ee} = \Sigma_{xx} - \Sigma_{xy}.\Sigma_{yy}^{-1}.\Sigma_{yx} \tag{2.23}$$

For any fixed values of **Y**, i.e. **y**, the vector variate **X** becomes the conditional vector variate **X(y)**, i.e. the value of **X** given the set of values **y**, and consequently (2.19) then reads,

$$\mathbf{X(y)} = \mu_x + \Sigma_{xy}.\Sigma_{yy}^{-1}.(\mathbf{y} - \mu_y) + \mathbf{e} \tag{2.24}$$

Thus, the conditional value of **X** given a fixed <u>linear</u> combination of realisations of **Y** will have mean $\bar{\mathbf{X}}(\mathbf{y})$ given by,

$$\bar{\mathbf{X}}(\mathbf{y}) = \mu_x + \Sigma_{xy}.\Sigma_{yy}^{-1}.(\mathbf{y} - \mu_y) \tag{2.25}$$

and covariance matrix Σ_{ee} given by (2.23), which gives exactly the same value as in (2.14), when Normality, rather than linearity was assumed.

2.6 FILTERING

In the filtering phase we relate the parameters of the filtered vector variate, $\mathbf{x}_{t-1}(t-1)$, to those of the filtered vector variate, $\mathbf{x}_t(t)$ using the results of section 2.5.

Firstly we utilise (2.10) to find the covariance of $y_t(t-1)$ and $\mathbf{x}_t(t-1)$.

$$\mathbb{C}ov[y_t(t-1),\ \mathbf{x}_t(t-1)^T] = \mathbb{C}ov[(\mathbf{v}^T.\mathbf{x}_t(t-1) + e_t(t-1)).\mathbf{x}_t(t-1)^T]$$

$$= \mathbb{C}ov[\mathbf{v}^T.\mathbf{x}_t(t-1).\mathbf{x}_t(t-1)^T] = \mathbf{v}^T.\mathbb{C}ov[\mathbf{x}_t(t-1).\mathbf{x}_t(t-1)^T] = \mathbf{v}^T.\Sigma_t(t-1)$$

....(2.26)

Hence combining (2.06), (2.11), (2.12) and (2.26), the joint, (unspecified), distribution of $\mathbf{x}_t(t-1)$ and $y_t(t-1)$ can be written as:

$$\begin{Bmatrix} \mathbf{x}_t(t-1) \\ y_t(t-1) \end{Bmatrix} \sim UD \left\{ \begin{matrix} \mu_t(t-1) \\ \mathbf{v}^T.\mu_t(t-1) \end{matrix} ; \begin{matrix} \Sigma_t(t-1) & \Sigma_t(t-1).\mathbf{v} \\ \mathbf{v}^T.\Sigma_t(t-1) & \mathbf{v}^T.\Sigma_t(t-1).\mathbf{v} + \sigma_e^2 \end{matrix} \right\}$$

....(2.27)

which is now in the partitioned form of (2.13).

Hence using the results of section 2.5, equations (2.22) and (2.24), the linearly conditional distribution of $\mathbf{x}_t(t-1)$ given the realisation of $y_t(t-1)$, which is simply y_t, is the distribution of the filtered value $\mathbf{x}_t(t)$, whose mean, $\mu_t(t)$ is given by:

$$\mu_t(t) = \mu_t(t-1) + \frac{(y_t - \mathbf{v}^T.\mu_t(t-1))}{(\mathbf{v}^T.\Sigma_t(t-1).\mathbf{v} + \sigma_e^2)} .\Sigma_t(t-1).\mathbf{v} \qquad (2.28)$$

and whose covariance matrix, $\Sigma_t(t)$, is given by:

$$\Sigma_t(t) = \Sigma_t(t-1) - \Sigma_t(t-1).\mathbf{v}.\mathbf{v}^T.\Sigma_t(t-1)/(\mathbf{v}^T.\Sigma_t(t-1).\mathbf{v} + \sigma_e^2) \qquad (2.29)$$

Combining equations (2.28) and (2.29), above, with the prediction equations in (2.08) and (2.09), we obtain the filtering equations which update the mean and covariance matrix of the filtered value $\mathbf{x}_{t-1}(t-1)$ to those of $\mathbf{x}_t(t)$. Note that since the bracketed term and the subscript are the same for filtered values, $\mu_t(t)$, $\mu_{t-1}(t-1)$, $\Sigma_t(t)$ and $\Sigma_{t-1}(t-1)$ have been abbreviated to μ_t, μ_{t-1}, Σ_t and Σ_{t-1}. Also $\mathbf{v}^T.\Theta$ which equals $(\vartheta_1, \vartheta_2, \ldots, \vartheta_m)$ is written simply as ϑ^T.

$$\mu_t = \Theta.\mu_{t-1} + \frac{(y_t - \vartheta^T.\mu_{t-1})}{(\vartheta^T.\Sigma_{t-1}.\vartheta + \sigma_a^2 + \sigma_e^2)}.(\Theta.\Sigma_{t-1}.\vartheta + \sigma_a^2.\mathbf{v}) \qquad (2.30)$$

$$\Sigma_t = \Theta.\Sigma_{t-1}.\Theta^T + \sigma_a^2.\mathbf{v}.\mathbf{v}^T - \frac{(\Theta.\Sigma_{t-1}.\vartheta + \sigma_a^2.\mathbf{v}).(\Theta.\Sigma_{t-1}.\vartheta + \sigma_a^2.\mathbf{v})^T}{(\vartheta^T.\Sigma_{t-1}.\vartheta + \sigma_a^2 + \sigma_e^2)} \qquad \ldots.(2.31)$$

The prediction and filtering equations (2.08), (2.09), (2.30) and (2.31) are collectively known as the Kalman filter after the control engineer who first derived them, (*Kalman, 1960*).

2.7 SMOOTHING

The derivation of smoothed or fixed interval estimates is not as straightforward as that for filtered estimates. Different approaches

can be found, although none of these prove to be particularly satisfactory. In (*Jazwinski,1970*), the proof relies on the invention of an artificial auxiliary variable. In (*Sage and Melsa,1971*), the result is only applicable if it can be assumed that the appropriate estimate is that which maximises a posterior density distribution. In (*Anderson and Moore,1979*), the proof takes up nearly a chapter during which a necessary visit is paid to derive fixed point estimates, which are not particularly relevant, and in (*Ansley and Kohn,1982*) the proof is somewhat of an acquired taste.

To maintain the consistency of these sections we shall demonstrate a proof which again utilises the results of section 2.5. The proof can be broken down into three stages.

In stage one, we obtain the joint, unspecified, distribution of the vectors $\mathbf{y}_T(t)$ and $\mathbf{x}_t(t)$, where $\mathbf{y}_T(t)$ has elements $y_{t+1}(t)$, $y_{t+2}(t)$, ... , $y_{T-1}(t)$, $y_T(t)$ and then use the results of section 2.5 to obtain the distribution of $\mathbf{x}_t(T)$.

Repeated application of (2.01) leads to the following expression for $\mathbf{x}_{t+k}$ in terms of $\mathbf{x}_t$ and structural errors a_{t+1} to a_{t+k},

$$\mathbf{x}_{t+k} = \Theta^k.\mathbf{x}_t + a_{t+1}.\Theta^{k-1}.\mathbf{v} + a_{t+2}.\Theta^{k-2}.\mathbf{v} + \ldots + a_{t+k}.\mathbf{v} \qquad (2.32)$$

Applying (2.03) to (2.32) for time point t+k then gives,

$$y_{t+k} = \mathbf{v}^T.\Theta^k.\mathbf{x}_t + a_{t+1}.\mathbf{v}^T.\Theta^{k-1}.\mathbf{v} + a_{t+2}.\mathbf{v}^T.\Theta^{k-2}.\mathbf{v} + \ldots + a_{t+k} + e_{t+k}$$

$$\ldots.(2.33)$$

Hence, taking expectations given values of y up to "t",

$$\mathbb{E}[y_{t+k}(t)] = \mathbf{v}^T.\Theta^k.\mathbb{E}[\mathbf{x}_t(t)] = \mathbf{v}^T.\Theta^k.\mu_t(t) \qquad (2.34)$$

$$\text{since,} \quad \mathbb{E}[a_{t+k}(t)] = \mathbb{E}[e_{t+k}(t)] = 0 \text{ for } k > 0 \qquad (2.35)$$

Here the conventional use of capital letters to denote random variables has been dropped, since they are distinguished from their realisations by having a bracketed term indicating the number of observations of y on which their distribution is based.

Note also that (2.33), and more to the point, (2.01) and (2.03) would hold given any set of values of y, however (2.34) and (2.35) would not and so it would make little sense to attempt proofs containing terms such as $a_t(t+k)$ since their expectations are far from obvious, and certainly not zero.

The situation is analogous to considering the probability of the fourth throw of a coin being a head, (a) when the fourth throw has yet to be made, and (b) when the fourth throw has been made and it is known that only one of the four resulted in a head. In (a) the probability is the obvious 1/2, whereas in (b) it is, the less than obvious, 1/4.

Also from (2.33), the covariance of $y_{t+k}(t)$ and $\mathbf{x}_t(t)$, for $k > 0$, is given by,

$$\mathbb{C}ov[y_{t+k}(t).\mathbf{x}_t(t)^T] = \mathbf{v}^T.\Theta^k.\mathbb{C}ov[\mathbf{x}_t(t).\mathbf{x}_t(t)^T] = \mathbf{v}^T.\Theta^k.\Sigma_t(t) \qquad (2.36)$$

since,

$$\mathbb{E}[a_{t+k}(t),\ \mathbf{x}_t(t)] = \mathbf{0} \quad \text{for} \quad k > 0 \qquad (2.37)$$

$$\mathbb{E}[e_{t+k}(t),\ \mathbf{x}_t(t)] = \mathbf{0} \quad \text{for} \quad k \geq 0 \qquad (2.38)$$

Hence, from (2.36), the $T-t \times d$ covariance matrix of the vectors $\mathbf{y}_{T-t}(t)$ and $\mathbf{x}_t(t)$, where $\mathbf{y}_{T-t}(t)$ has elements $y_{t+1}(t)$, $y_{t+2}(t)$, ... , $y_{T-1}(t)$, $y_T(t)$ is given by,

$$\mathbb{C}ov[\mathbf{y}_{T-t}(t).\mathbf{x}_t(t)^T] = = \Omega_1(\Theta).\Sigma_t(t) \qquad (2.39)$$

where $\Omega_1(\Theta)$ is the T-t × d matrix whose T-t rows are given by $\mathbf{v}^T.\Theta$, $\mathbf{v}^T.\Theta^2$, $\mathbf{v}^T.\Theta^3$, ... , $\mathbf{v}^T.\Theta^{T-t}$, (Not confusing $\mathbf{v}^T$, i.e. $\mathbf{v}$ transposed with Θ^{T-t} i.e. Θ to the power (T-t)).

And from (2.34) the mean of $\mathbf{y}_T(t)$ is given by,

$$\mathbb{E}[\mathbf{y}_{T-t}(t)] = \Omega_1(\Theta).\mu_t(t) \tag{2.40}$$

Combining the results of (2.36), (2.39) and (2.40), we obtain the joint, unspecified, distribution of the vectors $\mathbf{y}_{T-t}(t)$ and $\mathbf{x}_t(t)$, as:

$$\begin{pmatrix} \mathbf{x}_t(t) \\ \mathbf{y}_{T-t}(t) \end{pmatrix} \sim UD \left(\begin{array}{cc|c} \mu_t(t) & ; & \Sigma_t(t) \quad , \Sigma_t(t).\Omega_1(\Theta)^T \\ \Omega_1(\Theta).\mu_t(t) & ; & \Omega_1(\Theta).\Sigma_t(t), \quad \sum_{yy} \end{array} \right) \tag{2.41}$$

where $\sum_{yy}$ is the covariance matrix of $\mathbf{y}_{T-t}(t)$.

From section 2.5, the linearly conditional distribution of $\mathbf{x}_t(t)$ given $\mathbf{y}_{T-t}$ i.e. $\mathbf{x}_t(T)$ is given by:

$$\mathbf{x}_t(T) \sim UD \left(\mu_t(T) \; ; \; \Sigma_t(T) \right) \tag{2.42}$$

$$\text{where} \quad \mu_t(T) = \mu_t(t) + \Sigma_t(t).\Omega_1(\Theta)^T.\sum_{yy}^{-1}.\left(\mathbf{y}_{T-t} - \Omega_1(\Theta).\mu_t(t)\right) \tag{2.43}$$

$$\text{and} \quad \Sigma_t(T) = \Sigma_t(t) - \Sigma_t(t).\Omega_1(\Theta)^T.\sum_{yy}^{-1}.\Omega_1(\Theta).\Sigma_t(t) \tag{2.44}$$

In stage two, we use an almost identical argument to that of stage one to obtain the joint, unspecified, distribution of the vectors $\mathbf{y}_{T-t}(t)$ and $\mathbf{x}_{t+1}(t)$, where $\mathbf{y}_{T-t}(t)$ was defined in stage one, and then again use the results of section 2.3 to obtain the distribution of $\mathbf{x}_{t+1}(T)$.

In a similar way to obtaining (2.33), an expression for y_{t+k} in terms of $\mathbf{x}_{t+1}$ and structural errors a_{t+2} to a_{t+k} is given by,

$$y_{t+k} = \mathbf{v}^T.\Theta^{k-1}.\mathbf{x}_{t+1} + a_{t+2}.\mathbf{v}^T.\Theta^{k-2}.\mathbf{v} + \ldots + a_{t+k} + e_{t+k} \qquad (2.45)$$

Hence, reasoning as before, we obtain,

$$\mathbb{E}[y_{t+k}(t)] = \mathbf{v}^T.\Theta^k.\mathbb{E}[\mathbf{x}_{t+1}(t)] = \mathbf{v}^T.\Theta^{k-1}.\mu_{t+1}(t) \qquad (2.46)$$

$$\mathbb{C}ov[y_{t+k}(t).\mathbf{x}_{t+1}(t)^T] = \mathbf{v}^T\Theta^{k-1}.\mathbb{C}ov[\mathbf{x}_{t+1}(t).\mathbf{x}_{t+1}(t)^T] = \mathbf{v}^T.\Theta^{k-1}.\Sigma_{t+1}(t)$$

$$\ldots.(2.47)$$

and $$\mathbb{C}ov[\mathbf{y}_{T-t}(t).\mathbf{x}_{t+1}(t)^T] = \Omega_0(\Theta).\Sigma_{t+1}(t) \qquad (2.48)$$

where $\Omega_0(\Theta)$ is the T-t $\times$ d matrix whose T-t rows are given by $\mathbf{v}^T$, $\mathbf{v}^T.\Theta^1$, $\mathbf{v}^T.\Theta^2$, ... , $\mathbf{v}^T.\Theta^{T-t-1}$.

And from (2.46) the mean of $\mathbf{y}_{T-t}(t)$ is given by,

$$\mathbb{E}[\mathbf{y}_{T-t}(t)] = \Omega_0(\Theta).\mu_{t+1}(t) \qquad (2.49)$$

Combining the results of (2.47), (2.48) and (2.49), we obtain the joint, unspecified, distribution of the vectors $\mathbf{y}_{T-t}(t)$ and $\mathbf{x}_{t+1}(t)$, as:

$$\begin{pmatrix} \mathbf{x}_{t+1}(t) \\ \mathbf{y}_{T-t}(t) \end{pmatrix} \sim UD \begin{pmatrix} \mu_{t+1}(t) & ; & \Sigma_t(t) & , & \Sigma_{t+1}(t).\Omega_0(\Theta)^T \\ \Omega_0(\Theta).\mu_{t+1}(t) & ; & \Omega_0(\Theta).\Sigma_{t+1}(t), & & \sum_{yy} \end{pmatrix} \qquad (2.50)$$

where $\sum_{yy}$ is, as in stage one, the covariance matrix of $\mathbf{y}_{T-t}(t)$.

From section **2.5**, the linearly conditional distribution of $\mathbf{x}_{t+1}(t)$ given $\mathbf{y}_{T-t}$ i.e. $\mathbf{x}_{t+1}(T)$ is given by:

$$\mathbf{x}_{t+1}(T) \sim UD\left(\mu_{t+1}(T) \; ; \; \Sigma_{t+1}(T) \right) \tag{2.51}$$

with

$$\mu_{t+1}(T) = \mu_{t+1}(t) + \Sigma_{t+1}(t).\Omega_0(\Theta)^T.\sum_{yy}^{-1}.\left(\mathbf{y}_{T-t} - \Omega_0(\Theta).\mu_{t+1}(t)\right) \tag{2.52}$$

and $$\Sigma_{t+1}(T) = \Sigma_{t+1}(t) - \Sigma_{t+1}(t).\Omega_0(\Theta)^T.\sum_{yy}^{-1}.\Omega_0(\Theta).\Sigma_{t+1}(t) \tag{2.53}$$

Stage three combines the results of stages one and two by eliminating the covariance matrix of $\mathbf{y}_{T-t}(t)$, $\sum_{yy}$.

Rearranging (2.44) we have,

$$\Omega_1(\Theta)^T.\sum_{yy}^{-1}.\Omega_1(\Theta) = \Sigma_t^{-1}(t).(\Sigma_t(t)-\Sigma_t(T)).\Sigma_t^{-1}(t) \tag{2.54}$$

and rearranging (2.53) we similarly have,

$$\Omega_0(\Theta)^T.\sum_{yy}^{-1}.\Omega_0(\Theta) = \Sigma_{t+1}^{-1}(t).(\Sigma_{t+1}(t)-\Sigma_{t+1}(T)).\Sigma_{t+1}^{-1}(t) \tag{2.55}$$

Note also that from their definitions in (2.39) and (2.48),

$$\Omega_1(\Theta) = \Omega_0(\Theta).\Theta \tag{2.56}$$

Combining (2.54), (2.55) and (2.56) we get one of the two initial smoothing equations,

$$\boxed{\Sigma_t(T) = \Sigma_t(t) + \Sigma_t(t).\Theta^T.\Sigma_{t+1}^{-1}(t).\left(\Sigma_{t+1}(T)-\Sigma_{t+1}(t)\right).\Sigma_{t+1}^{-1}(t).\Theta.\Sigma_t(t)} \tag{2.57}$$

Finally manipulating equations (2.43) and (2.52) utilising (2.08) and (2.56), we obtain the other initial smoothing equation,

$$\mu_t(T) = \mu_t(t) + \Sigma_t(t).\Theta^T.\Sigma_{t+1}^{-1}(t).\left(\mu_{t+1}(T) - \mu_{t+1}(t) \right) \qquad (2.58)$$

which completes the proof, and the section on smoothing.

2.8 STARTING CONDITIONS

To initiate the processes of prediction, filtering and smoothing, we apparently need to know the values of $\mu_t(t)$ and $\Sigma_t(t)$, for some value of t, usually t=0 since then we require the values $\mu_0(0)$ and $\Sigma_0(0)$ i.e. the mean and variance of the distribution of $\mathbf{x}_0$, given in accordance with (2.02) as $\mathbf{x}_0 = (x_0, x_{-1}, \ldots, x_{-d+1})^T$, when no values of y_t have been observed at all. Put in Bayesian terms, we need to be able to specify its prior distribution, or for an unspecified distribution, its prior parameters.

How this is done has been the object of much controversy between statisticians. The Classical camp argue that to specify a prior distribution rather begs the question as to what the data will reveal and hence dismiss the Bayesian approach as being subjective and therefore necessarily biased.

The Bayesians, on the other hand, maintain that, even though a specified prior may not be completely accurate, it is almost always the case that some form of prior knowledge is available, and any attempt to utilise it in the form of the prior must therefore be better than no attempt at all.

The Bayesians quote the case of a coin being tossed and coming up heads four times in a row, which, by Classical, (Maximum Likelihood), reasoning, would suggest that the best estimate of its probability of falling heads is one, whereas, by choosing a common-sense prior

distribution whose mean is one half and whose variance is small, this dilemma is avoided, (see works by Lindley, 1965a, 1965b, 1971). The Classicists reply is that "life" is not as well understood as "coins". In practice, the confidence with which a prior can be specified will depend on the nature of the particular time series involved.

Let us suppose, for the moment that we have no knowledge whatsoever concerning $\mathbf{x}_0(0) = (x_0(0), x_{-1}(0), \ldots, x_{-d+1}(0))^T$. What does this imply as regards the choice of values for $\mu_0(0)$ and $\Sigma_0(0)$?

One obvious interpretation is that the variance of the prior distribution is infinite. If we genuinely have no information on $\mathbf{x}_0(0)$, then we must accept the equal possibility of it having any value, from minus infinity to plus infinity, implying a uniform distribution with infinite variance. Impractical as this may sound it does have some interesting consequences. It is known as a *vague* or *improper* prior, and, as a means to an end, which is what we need it for, it is perfectly acceptable. Lindley, (Lindley, 1965a, 1965b) shows that many of the results of Classical Inference are reproduced when Bayesian inference is applied with a *vague* prior.

Let us begin, therefore, with the assumption that the variances of each of the "d" variates, $x_0(0), x_{-1}(0), \ldots, x_{-d+1}(0)$, in the vector $\mathbf{x}_0(0)$ is infinite, i.e. that the diagonal elements of $\Sigma_0(0)$ are all ∞. What we are going to show is that, by successive applications of the filtering update equations, (2.29) and (2.30), the distribution of $\mathbf{x}_d(d) = (x_d(d), x_{d-1}(d), \ldots, x_1(d))^T$, has mean vector $\mu_d(d)$ given by $(y_d, y_{d-1}, \ldots, y_2, y_1)^T$ and covariance matrix $\Sigma_d(d)$ equal to $\sigma_e^2.\mathbf{I}_d$, where $\mathbf{I}_d$ is a dxd identity matrix, i.e. that the assumption of a vague prior is equivalent to assuming that each of the elements $x_i(d)$, i=1 to d, of $\mathbf{x}_d(d)$ is independently distributed with mean y_i and variance σ_e^2.

2.81 A NOTE ON COVARIANCE MATRICES WITH INFINITE VARIANCES

To do this we first need to establish a result for any strict, i.e. positive definite, covariance matrix, Σ.

Since Σ is positive definite then $\mathbf{z}^T.\Sigma.\mathbf{z} > 0$ for $\mathbf{z} \neq \emptyset$.

Suppose that we now partition the matrix Σ into matrices Σ_{11}, Σ_{12}, Σ_{21}, and Σ_{22}, where Σ_{11} contains the infinite diagonal variances and Σ_{22} does not, and $\mathbf{z}$ is partitioned into $\mathbf{z}_1$ and $\mathbf{z}_2$, likewise.

$$\text{Then, } \mathbf{z}^T.\Sigma.\mathbf{z} = \mathbf{z}_1^T.\Sigma_{11}.\mathbf{z}_1 + 2.\mathbf{z}_1^T.\Sigma_{12}.\mathbf{z}_2 + \mathbf{z}_2^T.\Sigma_{22}.\mathbf{z}_2 > 0 \qquad (2.59)$$

where, since Σ_{11} is also strictly positive definite, $\mathbf{z}_1^T.\Sigma_{11}.\mathbf{z}_1 > 0$ for $\mathbf{z}_1 \neq \emptyset$.

Letting each of the variances in Σ_{11} tend to σ_∞^2, (a very large value), whereby from (2.59), $\mathbf{z}^T.\Sigma.\mathbf{z} \rightarrow \sigma_\infty^2.\mathbf{z}_1^T.R_{11}.\mathbf{z}_1$, where R_{11} is the correlation matrix of Σ_{11}, i.e.

$$\mathbf{z}^T.\Sigma.\mathbf{z} \propto \sigma_\infty^2 \text{ for } \mathbf{z}_1 \neq \emptyset. \qquad (2.60)$$

Hence as $\sigma_\infty^2 \rightarrow \infty$, $\mathbf{z}^T.\Sigma.\mathbf{z} \rightarrow \infty$.

2.82 THE IMPLICATIONS OF ASSUMING VAGUE STARTING VALUES

By using the filtering update equation (2.31), the individual elements, $\sigma_{ij}(t)$, of the filtering covariance matrix, $\Sigma_t(t)$, and $\sigma_{ij}(t+1)$, of the filtering covariance matrix $\Sigma_{t+1}(t+1)$, are related by equations (2.61) to (2.64) below, namely,

$$\sigma_{11}(t+1) = \sigma_e^2.(\sigma^2 + \sigma_a^2)/(\sigma^2 + \sigma_a^2 + \sigma_e^2) \qquad (2.61)$$

$$\sigma_{i1}(t+1) = \sigma_e^2.\sigma_{i-1}(t).\bar{\sigma}_{i-1}/(\sigma^2 + \sigma_a^2 + \sigma_e^2) \text{ for } i > 1 \qquad (2.62)$$

$$\sigma_{ij}(t+1) = \sigma_{i-1,j-1}(t) - \sigma_{i-1}(t).\bar{\sigma}_{i-1}.\sigma_{j-1}(t).\bar{\sigma}_{j-1}/(\sigma^2 + \sigma_a^2 + \sigma_e^2)$$

$$\text{for } i,j > 1 \qquad (2.63)$$

$$\text{where } \bar{\sigma}_i = \sum^{j} \rho_{ij}(t).\sigma_j(t).\vartheta_j \text{ and } \sigma^2 = \vartheta^T.\Sigma.\vartheta = \sum^{i} \sigma_i(t).\bar{\sigma}_i.\vartheta_i \qquad (2.64)$$

Note that $\rho_{ij}(t)$ is a correlation, $\sigma_i(t)$ a standard deviation and ϑ_i a corresponding element of the vector ϑ, given in (2.02) and (2.31). Note also that any other undefined values are given by symmetry.

Hence, for example, for a 4×4 covariance matrix, Σ, whose original standard deviations, correlations and covariances were σ_i, ρ_{ij}, and σ_{ij}, the updated matrix would have the following 16 elements,

$$\begin{bmatrix}
\frac{\sigma_e^2.(\sigma^2+\sigma_a^2)}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \frac{\sigma_e^2.\sigma_1.\bar{\sigma}_1}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \frac{\sigma_e^2.\sigma_2.\bar{\sigma}_2}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \frac{\sigma_e^2.\sigma_3.\bar{\sigma}_3}{(\sigma^2+\sigma_a^2+\sigma_e^2)} \\
\frac{\sigma_e^2.\sigma_1.\bar{\sigma}_1}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{11}-\frac{\sigma_1.\bar{\sigma}_1.\sigma_1.\bar{\sigma}_1}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{12}-\frac{\sigma_1.\bar{\sigma}_1.\sigma_2.\bar{\sigma}_2}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{13}-\frac{\sigma_1.\bar{\sigma}_1.\sigma_3.\bar{\sigma}_3}{(\sigma^2+\sigma_a^2+\sigma_e^2)} \\
\frac{\sigma_e^2.\sigma_2.\bar{\sigma}_2}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{21}-\frac{\sigma_2.\bar{\sigma}_2.\sigma_1.\bar{\sigma}_1}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{22}-\frac{\sigma_2.\bar{\sigma}_2.\sigma_2.\bar{\sigma}_2}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{23}-\frac{\sigma_2.\bar{\sigma}_2.\sigma_3.\bar{\sigma}_3}{(\sigma^2+\sigma_a^2+\sigma_e^2)} \\
\frac{\sigma_e^2.\sigma_3.\bar{\sigma}_3}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{31}-\frac{\sigma_3.\bar{\sigma}_3.\sigma_1.\bar{\sigma}_1}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{32}-\frac{\sigma_3.\bar{\sigma}_3.\sigma_2.\bar{\sigma}_2}{(\sigma^2+\sigma_a^2+\sigma_e^2)}, & \sigma_{33}-\frac{\sigma_3.\bar{\sigma}_3.\sigma_3.\bar{\sigma}_3}{(\sigma^2+\sigma_a^2+\sigma_e^2)}
\end{bmatrix}$$

$$\text{where} \quad \bar{\sigma}_i = \sum_{j=1}^{j=4} \rho_{ij}.\sigma_j.\vartheta_j$$

$$\text{and} \quad \sigma^2 = \vartheta^T.\Sigma.\vartheta = \sum_{j=1}^{j=4}\sum_{i=1}^{i=4} \sigma_i.\sigma_j.\rho_{ij}.\vartheta_i.\vartheta_j = \sum_{i=1}^{i=4} \sigma_i.\bar{\sigma}_i.\vartheta_i$$

....(2.65)

Suppose we now begin with the vague prior covariance matrix $\Sigma_0(0)$ and apply the first filtering iteration to $\Sigma_1(1)$ using (2.65) as a guideline, letting the standard deviations of $\Sigma_0(0)$ tend to infinity via the very large value σ_∞.

$$\text{From (2.65)}, \ \bar{\sigma}_i \to \sigma_\infty.\sum^{j} \rho_{ij}.\vartheta_j \ \text{and} \ \sigma^2 \to \sigma_\infty^2.\sum^{i,j} \rho_{ij}.\vartheta_i.\vartheta_j \qquad (2.66)$$

$$\text{Hence,} \quad \sigma_{11}(1) = \sigma_e^2.(\sigma^2+\sigma_a^2)/(\sigma^2+\sigma_a^2+\sigma_e^2) \to \sigma_e^2,$$

and, $\sigma_{i1}(1) = \sigma_e^2.\sigma_{i-1}.\bar{\sigma}_{i-1}./(\sigma^2+\sigma_a^2+\sigma_e^2) \quad \rightarrow \sigma_e^2.\sum^{j} \rho_{i-1,j}.\vartheta_j \Big/ \sum^{i,j} \rho_{ij}.\vartheta_i.\vartheta_j$

for $i > 1$. Thus, as σ_∞ tends to infinity, the first row, (and first column), of the matrix $\Sigma_1(1)$ will have an initial element σ_e^2 and other elements which are all finite. It can also be seen that any other elements of $\Sigma_1(1)$ will still be infinite. In other words $\sigma_1(1)$ will tend to σ_e, but all other standard deviations will still tend to infinity, but with $\sigma_{i1}(1)$ finite which implies that for $i > 1$, all $\rho_{i1}(1)$ will tend to zero, since $\sigma_{i1}(1) = \sigma_i(1).\sigma_1(1).\rho_{i1}(1)$. Hence the first element of $\mathbf{x}_1(1)$, i.e. $x_1(1)$, will be independently distributed with variance σ_e^2.

Turning our attention to equation (2.30), we see that, the individual elements, $m_i(t)$, of the filtered mean vector, $\mu_t(t)$, and $m_i(t+1)$, of the filtered mean vector $\mu_{t+1}(t+1)$, are related by equations (2.67) to (2.69) below, namely,

$$m_1(t+1) = \bar{m} + (y_{t+1} - \bar{m}).(\sigma^2+\sigma_a^2)/(\sigma^2+\sigma_a^2+\sigma_e^2) \qquad (2.67)$$

$$m_i(t+1) = m_i(t) - (y_{t+1} - \bar{m}).\sigma_{i-1}(t).\bar{\sigma}_{i-1}/(\sigma^2 + \sigma_a^2 + \sigma_e^2) \text{ for } i > 1 \qquad \ldots.(2.68)$$

where $\bar{m} = \sum^{j} m_j(t).\vartheta_j$, with $\bar{\sigma}_i$ and σ^2 as defined in (2.64). (2.69)

Hence, for example, for a 4×1 vector, μ, whose original means were m_1, m_2, m_3, and m_4, associated with a covariance matrix Σ, as defined in (2.64), the updated, (using the observation y), mean vector would have elements,

$$\bar{m} - \frac{(\sigma^2+\sigma_a^2)}{(\sigma^2+\sigma_a^2+\sigma_e^2)}.(y-\bar{m})$$

$$m_1 - \frac{\sigma_1.\bar{\sigma}_1}{(\sigma^2+\sigma_a^2+\sigma_e^2)}.(y-\bar{m})$$

$$m_2 - \frac{\sigma_2.\bar{\sigma}_2}{(\sigma^2+\sigma_a^2+\sigma_e^2)}.(y-\bar{m})$$

$$m_3 - \frac{\sigma_3.\bar{\sigma}_3}{(\sigma^2+\sigma_a^2+\sigma_e^2)}.(y-\bar{m})$$

where $\bar{m} = \sum_{j=1}^{j=4} m_j.\vartheta_j$, with $\bar{\sigma}_i$ and σ^2 as defined in (2.64). (2.70)

If μ is $\mu_0(0)$, associated with the vague prior covariance matrix $\Sigma_0(0)$ and applying the first filtering iteration to $\mu_1(1)$ using (2.70) as a guideline, (letting the standard deviations of $\Sigma_0(0)$ tend to infinity via the very large value σ_∞). Then, from (2.66) as $\sigma_i \to \sigma_\infty$,

$$\bar{\sigma}_i \to \sigma_\infty.\sum^{j} \rho_{ij}.\vartheta_j \text{ and } \sigma^2 \to \sigma_\infty^2.\sum^{i,j} \rho_{ij}.\vartheta_i.\vartheta_j$$

Hence $(\sigma^2+\sigma_a^2)/(\sigma^2+\sigma_a^2+\sigma_e^2) \to 1$, and $\sigma_i.\bar{\sigma}_i/(\sigma^2+\sigma_a^2+\sigma_e^2)$ tend to finite values.

Therefore, from (2.70) and, the first element of $\mathbf{x}_1(1)$, i.e. $x_1(1)$, is not only independently distributed with variance σ_e^2 as previously shown, but also has mean y_1.

We next come to the second iteration, which moves us from filtered covariance matrix $\Sigma_1(1)$ to $\Sigma_2(2)$, and mean vector $\mu_1(1)$ to $\mu_2(2)$, and again we can use the updating procedures in (2.65) and (2.70) to guide us, remembering that our starting point now has $\sigma_1 = \sigma_e$, $m_1 = y_1$, $\rho_{1i} = 0$ and $\sigma_i = \sigma_\infty$, for i=2 to 4.

Again, $\bar{\sigma}_i \propto \sigma_\infty$, for i=2 to 3, and $\sigma^2 \propto \sigma_\infty^2$ from section **2.81**, but now $\bar{\sigma}_1$ is finite, since each σ_{1i} is finite. Inspection of (2.64), applying the same arguments as before, leads us to conclude that the filtering steps $\Sigma_0(0) \to \Sigma_1(1) \to \Sigma_2(2)$ have forms,

$$\begin{bmatrix} \infty & \infty & \infty & \infty \\ \infty & \infty & \infty & \infty \\ \infty & \infty & \infty & \infty \\ \infty & \infty & \infty & \infty \end{bmatrix} \rightarrow \begin{bmatrix} \sigma_e^2 & F & F & F \\ F & \infty & \infty & \infty \\ F & \infty & \infty & \infty \\ F & \infty & \infty & \infty \end{bmatrix} \rightarrow \begin{bmatrix} \sigma_e^2 & 0 & F & F \\ 0 & \sigma_e^2 & F & F \\ F & F & \infty & \infty \\ F & F & \infty & \infty \end{bmatrix}$$

where the elements F have finite values whose corresponding correlations are zero.

Continuing the iterations using the same arguments leads us to the final steps to $\Sigma_3(3)$ and to $\Sigma_4(4)$, i.e.

$$\rightarrow \begin{bmatrix} \sigma_e^2 & 0 & 0 & F \\ 0 & \sigma_e^2 & 0 & F \\ 0 & 0 & \sigma_e^2 & F \\ F & F & F & \infty \end{bmatrix} \rightarrow \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}$$

....(2.71)

Similarly, applying the same arguments to the mean vector $\mu_1(1)$ leads us to the complete series of iterations, $\mu_0(0) \rightarrow \mu_1(1) \rightarrow \mu_2(2) \rightarrow \mu_3(3) \rightarrow \mu_4(4)$, i.e.

$$\begin{bmatrix} ? \\ ? \\ ? \\ ? \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ ? \\ ? \\ ? \end{bmatrix} \rightarrow \begin{bmatrix} y_2 \\ y_1 \\ ? \\ ? \end{bmatrix} \rightarrow \begin{bmatrix} y_3 \\ y_2 \\ y_1 \\ ? \end{bmatrix} \rightarrow \begin{bmatrix} y_4 \\ y_3 \\ y_2 \\ y_1 \end{bmatrix}$$

Generalisation gives us the result that the assumption of infinite variances for the elements of the d×1 variate vector $\mathbf{x}_0(0)$ is exactly equivalent to assuming that each element $x_i(d)$ of the variate vector $\mathbf{x}_d(d)$ is independently distributed with mean y_i and variance σ_e^2.

Hence, in summary, if $\mathbf{x}_0(0)$ has a vague prior distribution, then

$$\mathbb{E}[\mathbf{x}_d(d)] = \mu_d(d) = \mathbf{y}_d \qquad (2.72)$$

$$\mathbb{C}ov[\mathbf{x}_d(d)] = \Sigma_d(d) = \sigma_e^2.\mathbf{I}_{T-d} \qquad (2.73)$$

Thus by choosing what seems at first to be the rather unwieldy values of infinity for the variances of $\Sigma_0(0)$, the covariance matrix of $\mathbf{x}_0(0)$, we are led to what appear to be quite sensible starting values for $\mu_d(d)$ and $\Sigma_d(d)$. Note that such a choice also relieves us of the need to specify a value for $\mu_0(0)$, and the correlation structure of $\Sigma_0(0)$.

2.9 MORE GENERAL MODELS

Whilst we have confined ourselves to the general autoregressive model in this chapter, it should be stressed that the proofs are easily extended to more general cases such as the one below described by measurement equation (2.74) and structural equation (2.75), (see section **8.2**, (and in particularly **8.22**), of chapter eight.

$$y_t = \mathbf{h}_t^T.\mathbf{x}_t + d_t + e_t \qquad (2.74)$$

$$\mathbf{x}_t = \mathbf{C}_t.\mathbf{x}_{t-1} + \mathbf{b}_t + \mathbf{H}_t.\mathbf{a}_t \qquad (2.75)$$

where $\mathbf{a}_t$ now becomes a vector of i.i.d. residuals, again with means zero and variances σ_a^2, and d_t, $\mathbf{h}_t$, $\mathbf{b}_t$, $\mathbf{C}_t$ and $\mathbf{H}_t$ are all <u>known</u> scalars, vectors and matrices which can vary with time.

THE STATE SPACE APPROACH TO TREND ESTIMATION II

3.1 A DIRECT DERIVATION OF STATE SPACE VALUES

The main purpose of this section is to demonstrate the equivalence of the State Space estimates of trend, (produced on the assumption of a vague prior) and the Classical, (GLS), estimates of chapter one.

In the last chapter we saw how successive values of the mean and variance matrix of the posterior d×1 vector $\mathbf{x}_t(T) = (x_t(T), x_{t-1}(T), \ldots, x_{t-d}(T))^T$ could be generated. We begin this chapter by showing how the same set of posterior parameters may be obtained directly, in one operation, thus obtaining the mean, $\mu_T(T)$, and covariance matrix, $\Sigma_T(T)$, of the total T×1 vector $\mathbf{x}_T(T) = (x_T(T), x_{T-1}(T), \ldots, x_1(T))^T$.

We begin with the starting values associated with vague prior knowledge as derived in section 2.8 of the last chapter, namely that the vector $\mathbf{x}_d(T) = (x_d(T), x_{d-1}(T), \ldots, x_1(T))^T$ should have mean vector $\mathbf{y}_d = (y_d, y_{d-1}, \ldots, y_1)^T$ and covariance matrix $\sigma_e^2.\mathbf{I}_d$, where $\mathbf{I}_d$ is the d×d identity matrix.

In chapter one, equation (1.46), we saw how the structural equations for the general autoregressive model could be written in matrix form as:

$$\mathbf{a}_{T-d} = \mathbf{D}.\mathbf{x}_T \tag{3.01}$$

$$\text{where,} \qquad \mathbf{a}_{T-d} = (a_T, a_{T-1}, \ldots, a_{d+1})^T \tag{3.02}$$

and **D** is the T-d × T difference matrix with structure:

$$\mathbf{D} = \begin{bmatrix} 1 & -\vartheta_1 & -\vartheta_2 & \ldots\ldots & -\vartheta_d & 0 & 0 & 0 & \ldots\ldots & 0 \\ 0 & 1 & -\vartheta_1 & -\vartheta_2 & \ldots\ldots & -\vartheta_d & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & & & \text{etc} \end{bmatrix} \tag{3.03}$$

We can partition the matrix **D** into two matrices **B** and **C**, i.e. **D** = [**B**¦-**C**], where **B** is a square T-d × T-d invertible matrix, as follows:

$$\mathbf{D} = [\mathbf{B}\,\vdots\,-\mathbf{C}] = \left[\begin{array}{cc} 1, -\vartheta_1, \ldots -\vartheta_d, 0, \ldots, 0 & 0, \ldots 0, 0, 0 \\ & \\ 0, \ldots\ldots\ldots, 0, 0, 1, -\vartheta_1 & -\vartheta_2, \ldots -\vartheta_{d-1}, 0 \\ 0, 0, \ldots, 0, 0, 0, 1 & -\vartheta_1, \ldots\ldots, -\vartheta_d \end{array}\right] \qquad (3.04)$$

and so we may write (3.01) as:

$$\mathbf{a}_{T-d} = \mathbf{D}.\mathbf{x}_T = \mathbf{B}.\mathbf{x}_{T-d} - \mathbf{C}.\mathbf{x}_d \qquad (3.05)$$

Rearranging (3.05) and conditioning on values y_1 to y_d, we get,

$$\mathbf{B}.\mathbf{x}_{T-d}(d) = \mathbf{C}.\mathbf{x}_d(d) + \mathbf{a}_{T-d}(d) \qquad (3.06)$$

where,

$$\begin{aligned} \mathbf{a}_{T-d}(d) &= (a_T(d), \ldots, a_{d+1}(d))^T \\ \mathbf{x}_{T-d}(d) &= (x_T(d), \ldots, x_{d+1}(d))^T \\ \mathbf{x}_d(d) &= (x_d(d), \ldots, x_1(d))^T \end{aligned} \qquad \ldots\ldots(3.07)$$

Hence from (3.06) we have,

$$\mathbf{x}_{T-d}(d) = \mathbf{B}^{-1}.\mathbf{C}.\mathbf{x}_d(d) + \mathbf{B}^{-1}.\mathbf{a}_{T-d}(d) \qquad (3.08)$$

Taking expectations of (3.08) gives,

$$\mu_{T-d}(d) = \mathbf{B}^{-1}.\mathbf{C}.\mu_d(d) = \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d \qquad (3.09)$$

since a vague prior implies $\mu_d(d) = \mathbf{y}_d$ from section **2.82**, equation (2.72), of the last chapter.

Similarly for covariances,

$$\Sigma_{T-d}(d) = B^{-1}.C.\Sigma_d(d).C^T(B^T)^{-1} + \sigma_a^2.B^{-1}.(B^T)^{-1}$$

$$= B^{-1}.\left[\ \sigma_e^2.C.C^T + \sigma_a^2.I_{T-d}\ \right].(B^T)^{-1} = B^{-1}.\Omega_{T-d}.(B^T)^{-1} - \sigma_e^2.I_{T-d}$$

$$\text{where} \quad \Omega_{T-d} = \sigma_a^2.I_{T-d} + \sigma_e^2.B.B^T + \sigma_e^2.C.C^T = \sigma_a^2.I_{T-d} + \sigma_e^2.D.D^T$$

....(3.10)

since again a vague prior implies $\Sigma_d(d) = \sigma_e^2.I_{T-d}$ from section 2.82, equation (2.73), of the last chapter.

Also, the covariance of $\mathbf{x}_{T-d}(d)$ and $\mathbf{x}_d(d)$ is obtained from (3.08) as,

$$\mathbb{C}ov[\mathbf{x}_{T-d}(d).\mathbf{x}_d(d)^T] = \sigma_e^2.B^{-1}.C \qquad (3.11)$$

The measurement equation for the final T-d measurements of the general autoregressive model, (see chapter one, (1.05), is, when similarly conditioned on y_1 to y_d,

$$\mathbf{y}_{T-d}(d) = \mathbf{x}_{T-d}(d) + e_{T-d}(d) \qquad (3.11)$$

where,

$$\mathbf{y}_{T-d}(d) = (y_T(d), \ \ldots\ , y_{d+1}(d))^T$$
$$\mathbf{e}_{T-d}(d) = (e_T(d), \ \ldots\ , e_{d+1}(d))^T$$

....(3.12)

From (3.11) the mean vector and variance matrix of $\mathbf{y}_{T-d}(d)$, and the covariance matrix of $\mathbf{y}_{T-d}(d)$ and $\mathbf{x}_{T-d}(d)$ are given as:

$$\mathbb{E}[\mathbf{y}_{T-d}(d)] = \mathbb{E}[\mathbf{x}_{T-d}(d)] = \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d \qquad (3.13)$$

$$\mathbb{C}ov[\mathbf{y}_{T-d}(d)] = \Sigma_{T-d}(d) + \sigma_e^2.\mathbf{I}_{T-d} = \mathbf{B}^{-1}.\Omega_{T-d}.(\mathbf{B}^T)^{-1} \qquad (3.14)$$

$$\mathbb{C}ov[\mathbf{y}_{T-d}(d).\mathbf{x}_{T-d}(d)^T] = \Sigma_{T-d}(d) = \mathbf{B}^{-1}.\Omega_{T-d}.(\mathbf{B}^T)^{-1} - \sigma_e^2.\mathbf{I}_{T-d} \qquad (3.15)$$

Finally, combining (3.08) and (3.11), we can write,

$$\mathbf{y}_{T-d}(d) = \mathbf{B}^{-1}.\mathbf{C}.\mathbf{x}_d(d) + \mathbf{B}^{-1}.\mathbf{a}_{T-d}(d) + \mathbf{e}_{T-d}(d) \qquad (3.16)$$

from which the covariance of $\mathbf{y}_{T-d}(d)$ and $\mathbf{x}_d(d)$ is,

$$\mathbb{C}ov[\mathbf{y}_{T-d}(d).\mathbf{x}_d(d)^T] = \sigma_e^2.\mathbf{B}^{-1}.\mathbf{C} \qquad (3.17)$$

Equations (3.09) to (3.17), above provide us with the mean vectors and covariance matrices to be able to write the joint distribution of $\mathbf{y}_{T-d}(d)$, $\mathbf{x}_{T-d}(d)$ and $\mathbf{x}_d(d)$, i.e. $\mathbf{y}_{T-d}(d)$ and $\mathbf{x}_T(d)$, as

the vector $\begin{bmatrix} \mathbf{y}_{T-d}(d) \\ \mathbf{x}_{T-d}(d) \\ \mathbf{x}_d(d) \end{bmatrix}$ having mean vector $\begin{bmatrix} \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d \\ \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d \\ \mathbf{y}_d \end{bmatrix}$ and covariance

$$\text{matrix} \begin{bmatrix} (\mathbf{B}^T.\Omega_{T-d}^{-1}.\mathbf{B})^{-1} & (\mathbf{B}^T.\Omega_{T-d}^{-1}.\mathbf{B})^{-1}-\sigma_e^2.\mathbf{I}_{T-d} & \sigma_e^2.\mathbf{B}^{-1}.\mathbf{C} \\ (\mathbf{B}^T.\Omega_{T-d}^{-1}.\mathbf{B})^{-1}-\sigma_e^2.\mathbf{I}_{T-d} & (\mathbf{B}^T.\Omega_{T-d}^{-1}.\mathbf{B})^{-1}-\sigma_e^2.\mathbf{I}_{T-d} & \sigma_e^2.\mathbf{B}^{-1}.\mathbf{C} \\ \sigma_e^2.\mathbf{C}^T.(\mathbf{B}^T)^{-1} & \sigma_e^2.\mathbf{C}^T.(\mathbf{B}^T)^{-1} & \sigma_e^2.\mathbf{I}_d \end{bmatrix}$$

....(3.18)

We can now apply the results of section 2.5 to (3.18) to obtain the conditional mean vector, $\mu_T(T)$, and covariance matrix, $\Sigma_T(T)$, of $\mathbf{x}_T(T) = (\mathbf{x}_{T-d}(T), \mathbf{x}_d(T))^T$ as:

$$\begin{matrix}\mathbf{x}_{T-d}(T) \\ \mathbf{x}_d(T)\end{matrix} \sim UD \begin{bmatrix} \mathbf{y}_{T-d} - \sigma_e^2.\mathbf{B}^T\Omega_{T-d}^{-1}(\mathbf{B}.\mathbf{y}_{T-d} - \mathbf{C}.\mathbf{y}_d); & \sigma_e^2 - \sigma_e^4.\mathbf{B}^T\Omega_{T-d}^{-1}.\mathbf{B}, & \sigma_e^4.\mathbf{B}^T\Omega_{T-d}^{-1}.\mathbf{C} \\ \mathbf{y}_d + \sigma_e^2.\mathbf{C}^T\Omega_{T-d}^{-1}(\mathbf{B}.\mathbf{y}_{T-d} - \mathbf{C}.\mathbf{y}_d); & \sigma_e^4.\mathbf{C}^T\Omega_{T-d}^{-1}.\mathbf{B}, & \sigma_e^2 - \sigma_e^4.\mathbf{C}^T\Omega_{T-d}^{-1}.\mathbf{C} \end{bmatrix}$$

....(3.19)

which simplifies, using (3.04), to

$$\mathbf{x}_T(T) \sim UD\left[(\mathbf{I}_T - \sigma_e^2.\mathbf{D}^T.\Omega_{T-d}^{-1}.\mathbf{D}).\mathbf{y}_T;\ \sigma_e^2.(\mathbf{I}_T - \sigma_e^2.\mathbf{D}^T.\Omega_{T-d}^{-1}.\mathbf{D}) \right] \qquad (3.20)$$

Further, direct multiplication, using (3.10), confirms the relations,

$$\mathbf{I}_T - \sigma_e^2.\mathbf{D}^T.\Omega_{T-d}^{-1}.\mathbf{D} = (\mathbf{I}_T + \sigma_e^2/\sigma_a^2.\mathbf{D}^T.\mathbf{D})^{-1} = (\mathbf{I}_T + 1/\omega.\mathbf{D}^T.\mathbf{D})^{-1} \qquad (3.21)$$

and hence (3.20) becomes, (noting $\omega = \sigma_a^2/\sigma_e^2$),

$$\mathbf{x}_T(T) \sim UD\left[\mu_T(T);\Sigma_T(T)\right] \sim UD\left[(\mathbf{I}_T+1/\omega.\mathbf{D}^T.\mathbf{D})^{-1}.\mathbf{y}_T;\ \sigma_e^2.(\mathbf{I}_T+1/\omega.\mathbf{D}^T.\mathbf{D})^{-1}\right]$$

....(3.22)

which are identical to the Classical, (GLS), estimates of chapter one, equations (1.37) and (1.38) and to Whittaker's estimate of (1.14).

3.2 THE POSTERIOR MEAN AS AN ESTIMATOR

In the last section we derived the values of the mean vector, $\mu_T(T)$, and covariance matrix, $\Sigma_T(T)$, of the posterior distribution of the trend vector $\mathbf{x}_T$ given the data points $y_1, y_2, \ldots, y_T$, (more concisely $\mathbf{y}_T$), i.e. $\mathbf{x}_T(T)$.

Hence, for an individual element of $\mathbf{x}_T(T)$, say $x_t(T)$, its mean is given by the corresponding element of $\mu_T(T)$, say $\mu_t(T)$, which, for this section, overrides the last section's definition of $\mu_t(T)$, which was a vector, and its variance by the corresponding diagonal element of $\Sigma_T(T)$, $\sigma_t^2(T)$.

Unless Normality is assumed, by placing it on the residual errors e_t and a_t, and also on the form of the prior distribution, the values of the mean and variance are all we know about $x_t(T)$. The question now arises as to whether this is sufficient to produce an optimal estimate of $x_t(T)$ when its distribution is unknown.

Since we are free to choose any value, say ξ, of $x_t(T)$, on the basis of minimising its mean squared error, then ξ will be the value which minimises,

$$\mathbb{E}\left[(\xi - x_t(T))^2 \right] \tag{3.23}$$

over all values of $x_t(T)$.

Differentiating with respect to ξ and setting to zero gives,

$$\mathbb{E}[\xi] = \mathbb{E}[x_t(T)] \tag{3.24}$$

or in integral form,

$$\int \xi . p(x_t(T)) . dx_t(T) = \int x_t(T) . p(x_t(T)) . dx_t(T) \tag{3.25}$$

Since ξ is not a function of $x_t(T)$ and the right hand side of the expression is just the mean of the distribution, we have:

$$\xi = \mu_t(T) \tag{3.26}$$

i.e. the estimate with minimum mean square error is just the mean of the posterior distribution.

The value $\xi=\mu_t(T)$ is known as a fixed interval or smoothed estimate, this term having been first used by control engineers.

In a similar way, the mean, $\mu_t(t)$, of the filtered variate, $x_t(t)$, can be shown to be optimal for the estimation of $x_t(t)$ and hence is known as a filtered estimate.

Finally, note that when the mean of the posterior distribution of the trend value is chosen as its estimator, then, because of (3.23), the variance of the distribution becomes the mean squared error of the estimator, thus conveniently utilising the only two pieces of information known about the distribution.

3.3 THE BASIC MODEL OF WHITTAKER'S PROBLEM

In chapter one we looked at the "Basic" model associated with Whittaker's problem The State Space equivalent of this is produced by setting d=1 and $\vartheta_1=1$ in equations (2.01), (2.02) and (2.03) of chapter two. The measurement equation then becomes:

$$y_t = x_t + e_t \tag{3.27}$$

and the associated structural equation is:

$$x_t = x_{t-1} + a_t \tag{3.28}$$

with the usual independence assumptions, (chapter two, (2.04)), placed on the e_t and a_t.

3.31 STARTING CONDITIONS

Assuming a vague prior distribution for x_0, i.e. that the variance of $x_0(0) = \sigma_0^2(0) = \infty$, gives *proper* starting values for the mean and variance of $x_1(1)$, in accordance with section **2.82** of chapter two, of

$\mu_1(1)$ and $\sigma_1^2(1)$, where:

$$\mu_1(1) = y_1 \quad \text{and} \quad \sigma_1^2(1) = \sigma_e^2 \tag{3.29}$$

3.32 PREDICTION

For the basic model, the prediction equations, (2.08) and (2.09), of chapter two become, (bearing in mind that $\mu_t(t-1)$ and $\mu_{t-1}(t-1)$ are now scalar),

$$\boxed{\mu_t(t-1) = \mu_{t-1}(t-1)} \tag{3.30}$$

and

$$\boxed{\sigma_t^2(t-1) = \sigma_{t-1}^2(t-1) + \sigma_a^2} \tag{3.31}$$

3.33 FILTERING

The filtering equations, (2.29) and (2.30), of chapter two become,

$$\boxed{\mu_t(t) = \alpha_t.\mu_{t-1}(t-1) + (1-\alpha_t).y_t} \tag{3.32}$$

and

$$\boxed{\sigma_t^2(t) = (1 - \alpha_t).\sigma_e^2} \tag{3.33}$$

$$\text{where} \quad \alpha_t = \sigma_e^2/(\sigma_{t-1}^2(t-1) + \sigma_a^2 + \sigma_e^2) \tag{3.34}$$

The term $k_t = 1-\alpha_t$, in (2.18) and (2.19) is often used in control engineering. It is known as the gain. For this model, because of (3.34), any α_t, and hence any k_t, must lie between between 0 and 1.

For t=1 either (3.33) or (3.34) gives $\alpha_1=0$, since $\sigma_0^2(0) = \infty$ and $\sigma_1^2(1) = \sigma_e^2$. Similarly substitution of (3.33), at t=t-1, into (3.34) gives us the generating mechanism for each successive values of α_t as:

$$\alpha_t = \sigma_e^2 / \left(\sigma_a^2 + 2.\sigma_e^2 - \alpha_{t-1}.\sigma_e^2 \right) \tag{3.35}$$

which, on defining the noise variance ratio, ω, in the usual way, as,

$$\omega = \sigma_a^2 / \sigma_e^2 \tag{3.36}$$

simplifies (3.35) to:

$$\alpha_t = 1 / \left(\omega + 2 - \alpha_{t-1} \right) \tag{3.37}$$

In (3.37), as $t \to \infty$, α_t and $\alpha_{t-1} \to \alpha_\infty$, since α_t is strictly monotonic. Hence,

$$\alpha_\infty = 1 / \left(\omega + 2 - \alpha_\infty \right) \tag{3.38}$$

$$\text{i.e. } \alpha_\infty^2 - (2+\omega)\alpha_\infty + 1 = 0 \tag{3.39}$$

We have met this form of equation before, in chapter one, (1.18). The root lying between 0 and 1, which α_∞ must do because of (3.34), is given by:

$$\boxed{\alpha_\infty = 1+\omega/2 - \sqrt{(1+\omega/2)^2-1}} \tag{3.40}$$

This equation is identical to that of (1.20) with α_∞ replacing λ, and ω replacing w. Also notice the similarity between (3.32) and an equation we met in chapter one, (1.24). As t tends to infinity (3.32) tends to (1.24), since α_t tends to λ.

Alpha values for different w ratios
Figure 3.1

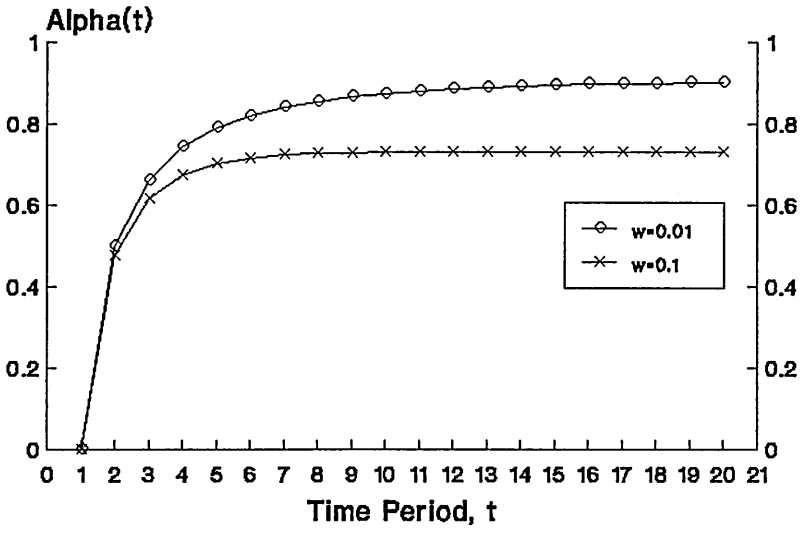


Figure 3.1 shows the values of α_t for two different values of ω; (We have chosen a starting value of $\alpha_1 = 0$ inferring a vague prior). Note that convergence is fairly rapid, being faster, the larger the value of ω. In terms of equation (3.32), this implies that for largish values of t, the process of filtering becomes that of simple exponential smoothing.

3.34 SMOOTHING

The initial smoothing equations, (2.56) and (2.57), of chapter two can be combined with the prediction equations to become,

$$\mu_t(T) = \beta_{t+1}.\mu_{t+1}(T) + (1-\beta_{t+1}).\mu_t(t) \qquad (3.41)$$

and

$$\sigma_t^2(T) = \beta_{t+1}^2 . \sigma_{t+1}^2(T) + (1-\beta_{t+1}) . \sigma_t^2(t) \qquad (3.42)$$

$$\text{where } \beta_{t+1} = \sigma_t^2(t)/(\sigma_t^2(t) + \sigma_a^2) \qquad (3.43)$$

Note that the processes in (3.41) and (3.42) operate backwards in time, beginning with the filtered values, $\mu_T(T)$ and $\sigma_T^2(T)$.

Note also that because of (3.43) and (3.33), the coefficients α_t and β_t are related by,

$$\beta_{t+1} = (1-\alpha_t) . \sigma_e^2 / \left((1-\alpha_t) . \sigma_e^2 + \sigma_a^2 \right) \qquad (3.44)$$

which when substituted in (3.37) gives the generating mechanism for β_t, namely,

$$\beta_{t+1} = 1 / \left(\omega + 2 - \beta_t \right) \qquad (3.45)$$

which is exactly the same mechanism as the one for α_t in (3.37), which in turn, in the light of (3.38), (3.39) and (3.40), together with the fact that any β_t must lie between 0 and 1 because of (3.43), means that:

$$\beta_\infty = 1+\omega/2 - \sqrt{(1+\omega/2)^2-1} \qquad (3.46)$$

Again, (3.46) above, has exactly the same form as (1.20) in chapter one, with β_∞ replacing λ.

The thing that distinguishes α_t and β_t is their starting values. For t=0 (3.43) gives β_1=1, since $\sigma_0^2(0) = \infty$, whereas as we saw α_1 was zero.

Beta values for different w ratios
Figure 3.2

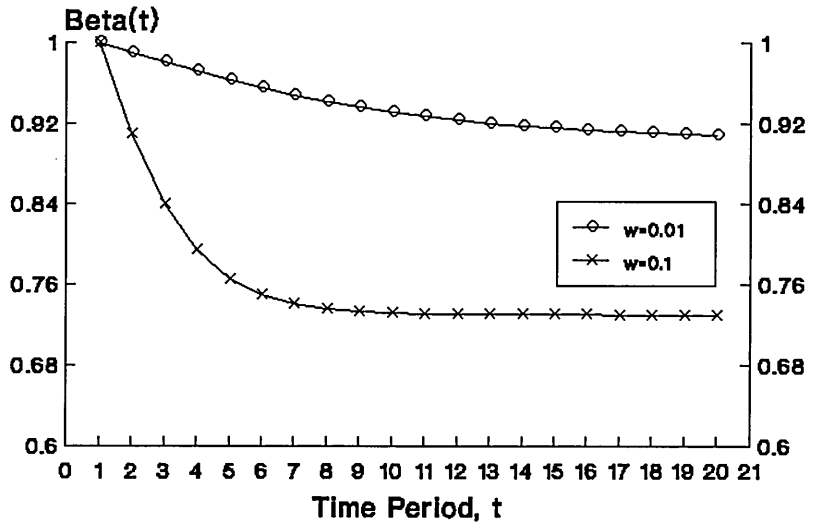


Figure 3.2 shows the variation in the values of β_t for two different values of w, the variance ratio. Like α_t, which can be regarded as its filtered equivalent, convergence is quite rapid, the more so for higher w. As β_t converges, equation (3.41) reduces to simple exponential smoothing applied backwards to the filtered series $\mu_t(t)$.

Finally by manipulating (3.44) and (3.45) we have:

$$\alpha_t = (1 - \beta_t) \; / \; (1 - \beta_t + \omega) \qquad (3.47)$$

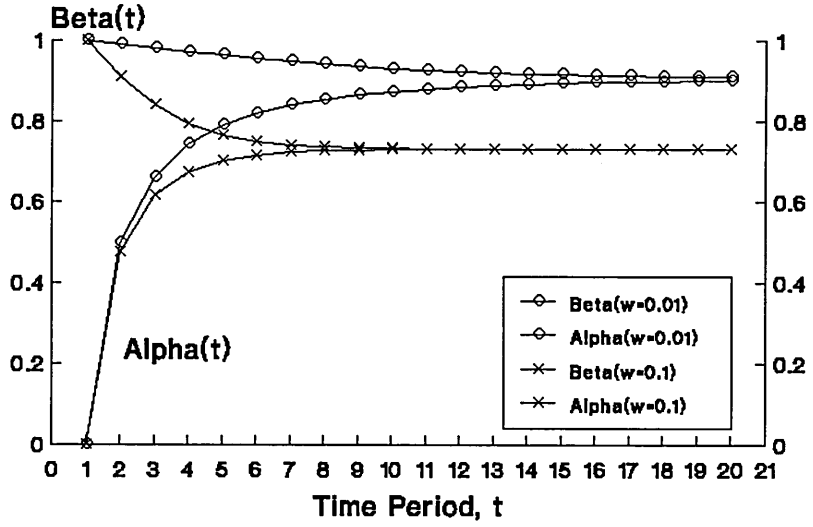
or alternatively:

$$\beta_t = (1 - \alpha_t - \omega.\alpha_t) \; / \; (1 - \alpha_t) \qquad (3.48)$$

Figure 3.3 shows the relationship between the coefficients α_t and β_t, demonstrating their convergence to a common value, which increases as

ω, the variance ratio is reduced.

Covergence of Alpha and Beta values
Figure 3.3



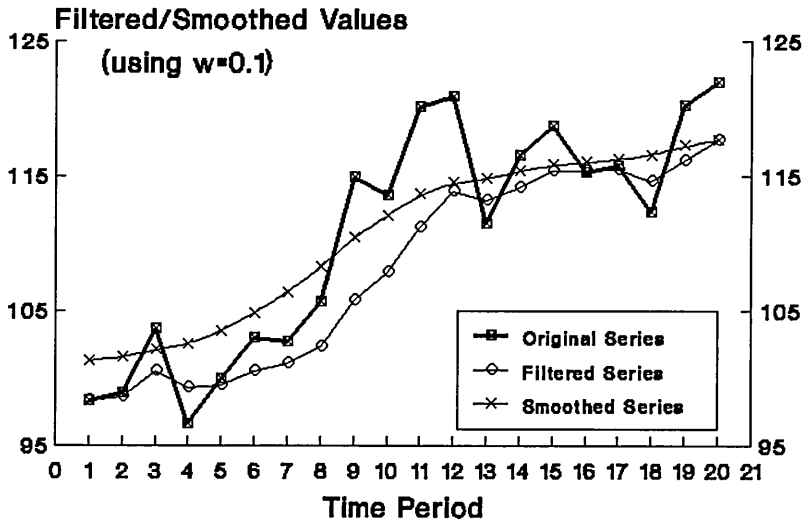
Note also that the larger the value of α_t, the smaller the value of β_t. In other words if the α_t are relatively large in the filtering equation for trend, i.e. (3.32) it suggests that the latest filtered value $\mu_t(t)$ is highly dependent on the previous value $\mu_{t-1}(t-1)$; whereas the reverse procedure in the smoothing equation, (3.41) would then necessarily incorporate relatively small values of β_t, implying that the smoothed value, $\mu_t(T)$, was also highly dependent on its filtered equivalent $\mu_t(t)$.

The outcome of this is that small values of ω lead to relatively large values of α_t and small values of β_t, which in turn produce both filtered and smoothed series which are dominated by initial filtered values i.e. are quite smooth.

By considering the alternative case, we can see a small variance ratio, ω, results in filtered and smoothed estimates which are both highly dependent on the original series y_t and therefore tend to follow the original series quite closely.

Figure 3.4 demonstrates the two processes of filtering and smoothing, the filtered values being produced using equation (3.32) and the smoothed values using equation (3.41).

Filtering and Smoothing
Figure 3.4



The filtered values can be thought of as an intermediate stage in the production of the smoothed values. Remember, they, the $\mu_t(t)$, are each calculated using only data values y_t up to time t, whereas the smoothed values, $\mu_t(T)$, utilise all data values up to y_T. Roughly, filtering can be thought of as a process of <u>forward</u> filtering, whereas smoothing is produced by further <u>backward</u> filtering the forward estimates.

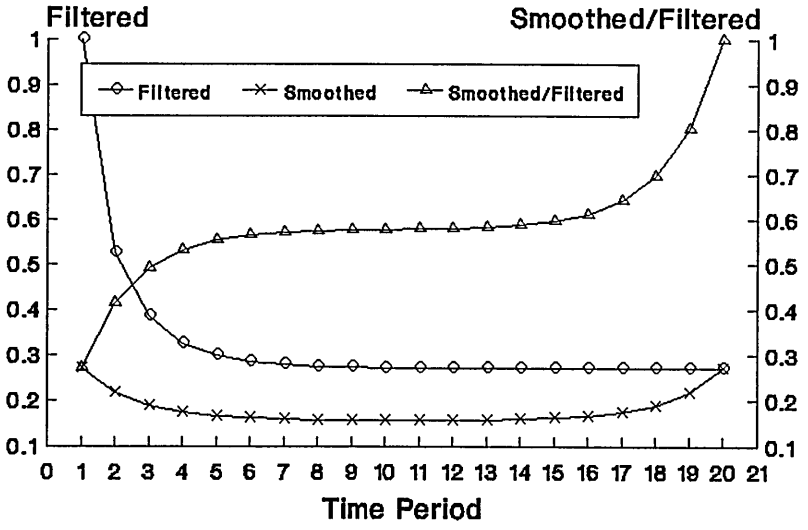
We may also note, as proved, that these smoothed values are identical to Whittaker's estimates of figure 1.1 in the last chapter.

3.35 FILTERED AND SMOOTHED VARIANCES

Figure 3.5 compares the variances, $\sigma_t^2(t)$ and $\sigma_t^2(T)$, of the distributions of the filtered, $x_t(t)$, and smoothed, $x_t(T)$, values, generated from equations (3.33) and (3.42) and standardised, (by dividing by σ_e^2).

From section 3.2, we note that these two variances also become the mean squared error of the respective trend estimates, when their mean are used as estimators.

Standardised Variances (w=0.1)
Figure 3.5



As one would expect, the filtered variance decreases as more data is gathered, however the smoothed variance does not behave quite so intuitively as maybe would be first expected, reaching a minimum value halfway through the series. The fact that this occurs halfway through the series, or that it regains its initial value exactly at the end of the series are not the points at issue here since it can be demonstrated that these effects do not arise in the general case. However the reaching a minimum does, and as can be demonstrated from the equations always does, whatever values of σ_e^2 and σ_a^2 are used.

The answer lies in the nature of the smoothing process itself. It is not the total amount of data we have available that determines how confident we would feel about a smoothed value, but rather how much data actually *surrounds* the trend value we are interested in. The more data we have coming both before and after the time period we are interested in, the more sure we will be of placing those events, (in this case the trend), in their proper context.

As any historian will tell you, hindsight is wonderful thing, but ask him to go too far back in time and things become very blurred, since his knowledge of what events preceded those in question are themselves very vague, (*he has a vague prior*). Hence what first seems to be a model inadequacy is on closer inspection just the opposite, and in a most satisfactory way is seen to be one of its strengths.

There is also another odd irony to this when spotted, and that is that the equations used to generate the filtered and smoothed variances do not depend on the actual values of y_t that have been observed, but only on their number. One could say that they <u>do</u> depend on the values of σ_e^2 and σ_a^2 used, which of necessity will require the actual values of y_t to be estimated, but this misses the point, namely that in theory we may choose $T=\infty$, or at least very large, to generate the variances, which would lead to filtered variances of zero, whatever values of σ_e^2 and σ_a^2 are used, as long as they are finite.

Again the model has a rather satisfactory solution. Yes, it is true that in theory smoothed variances of zero can be generated for any trend value, however the equations which generate the trend values, or more correctly the means of their distributions, <u>do</u> require the values of y_t. We can know the variance of the distribution of any x_t to any accuracy we like, but without the data, we would have no idea as to its mean. It would appear that the model incorporates its own version of Heisenberg's Uncertainty Principle.

Also plotted on the same graph graph is the value of the smoothing/filtering variance ratio $\sigma_t^2(T)/\sigma_t^2(t)$, which behaves as one would expect, having a value of one at t=T, and consistently falling the further back in time we go, demonstrating that it is the earlier values of t which benefit most, since it is they that obtain the most extra data from the smoothing process.

3.4 FURTHER INTERPRETATION OF THE STRUCTURAL EQUATION

Let us consider, what might be thought of as the second simplest

structural trend model to the basic model so far described in this chapter, namely,

$$x_t = 2.x_{t-1} - x_{t-2} + a_t \qquad (3.49)$$

This model would certainly satisfy the limitation of linear invariance described in section 1.3 of chapter one, since a linear trend such as (3.50) below,

$$x_t = a + b.t \qquad (3.50)$$

would pass through (3.49) undisturbed, since from (3.49) and (3.50),

$$a_t = x_t - 2.x_{t-1} + x_{t-2} = a+bt - 2(a+b(t-1)) + a+b(t-2) = 0 \qquad (3.51)$$

In terms of state space modelling, the equivalent state space structural equation would be, using the format of (2.01) of chapter two, i.e.

$$\mathbf{x}_t = \Theta.\mathbf{x}_{t-1} + a_t.\mathbf{v} \qquad (3.52)$$

given by:

$$\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} . \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix} + a_t . \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad (3.53)$$

where, what is known as the state vector, $\mathbf{x}_t$, comprises the elements x_t and x_{t-1}.

However, a state space modeller would almost certainly not use this structural formulation. Instead, he/she would write the model in

(3.49) as a set of two structural equations, namely (3.54) and (3.55) where,

$$x_t = x_{t-1} + g_t \tag{3.53}$$

$$g_t = g_{t-1} + a_t \tag{3.54}$$

and he/she would interpret the g_t as the gradient. His/her formulation would become,

$$\begin{pmatrix} x_t \\ g_t \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} . \begin{pmatrix} x_{t-1} \\ g_{t-1} \end{pmatrix} + a_t . \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3.55}$$

where the state vector $\mathbf{x}_t$, now consisted of the elements x_t and g_t, the trend value and the gradient at time t. His/her philosophy in doing so, is that information on the "state" of the system at any time t is contained in the state vector consisting of the trend and gradient.

However, we can see that the trend/gradient model formulation of (3.55) contains no more information than the general trend formulation of (3.53), nor is there any saving in complexity, since both models need to utilise state vectors and matrices of size two. Indeed, any information that one formulation produces can equally be gained by manipulation of the other.

Since this thesis concentrates on "trend estimation", it is advantageous for us to formulate everything into a general trend model, but this does not mean that we are overlooking the effects of gradient or curvature or higher order effects, just that we do not need to focus on them specifically. Indeed, we shall show that state space formulations of the type used in (3.55) can, at best, only equal the information content of the general autoregressive model and at worst are only a strict subset of it.

3.5 SUMMARY OF TREND ESTIMATION RESULTS

So far we have looked at three "optimal" techniques for obtaining trend estimates. In chapter one we considered the minimisation of Whittaker's function and also the Classical approach of Generalised Least Squares, (GLS). These were shown to produce identical values if we set a weighting factor, ω, in Whittaker's function equal to the ratio of the residual variances, σ_a^2/σ_e^2, of the model on which the procedure of Generalised Least Squares was applied.

In chapter two we investigated trend estimation using a State Space formulation of the model of chapter one, and concluded, at the beginning of this chapter by proving that we would again produce identical values for the trend estimates as previously if it was assumed that initial trend variates had infinite variances, i.e. a vague prior distribution, (we also showed that, given this assumption, the mean squared "errors" associated with the State Space and GLS estimates were also the same - Whittaker's approach, being algorithmic, had no equivalent concept).

However, this is far from the end of the story, since implicit in the trend estimates formulae, were, the residual variances, (or ω), and, for the variable parameter case, the autoregressive parameters. Hence, to complete the estimation procedure, we still require estimates for both these.

In the next chapter, and chapter five, we address the first of the above two areas, namely the estimation of residual variances.

ESTIMATION OF RESIDUAL VARIANCES I

All the work done so far implicitly assumed that we had knowledge of the measurement and structural variances, σ_a^2 and σ_e^2, and/or their ratio, ω.

In most situations we have no such knowledge, and hence need to estimate them from the only available data, namely the T×1 vector, $\mathbf{y}_T$, of observations, y_T, y_{T-1}, ... , and y_1.

In both this chapter and the next we consider the different ways in which this could be done.

4.1 THE LOG-LIKELIHOOD FUNCTION

If the density of the observed data is known in terms of the unknown parameters, then a well-established approach to the estimation of the unknown parameters, because of its desirable properties, is to choose those parameter values which maximise the probability density of the observed data, p(data/parameters) also known as the likelihood function of the parameters, i.e $\mathbb{L}$(parameters), or equivalently its logarithm, the log-likelihood function, $\mathbb{LL}$(parameters).

4.11 THE ASSUMPTION OF NORMALITY

Because of stochastic Normality assumptions and model linearities, many derived data vector distributions turn out to be multivariate Normal, with mean vector μ and covariance matrix Σ, i.e. the relevant T×1 data vector $\mathbf{y} \sim N(\vartheta,\Sigma)$. Under these conditions the log-likelihood function, $\mathbb{LL}(\mu,\Sigma)$, of the unknown parameters in μ and Σ is given by,

$$\ln(p(\mathbf{y}/\mu,\Sigma)) = \mathbb{LL}(\mu,\Sigma) = -1/2.(T.\ln(2\pi) + \ln|\Sigma| + (\mathbf{y}-\mu)^T.\Sigma^{-1}.(\mathbf{y}-\mu))$$

where $|\Sigma|$ is the determinant of Σ (4.01)

4.2 THE STATE SPACE APPROACH

As we have seen in chapter two, section 2.8, the State Space approach to trend estimation required us to specify some form of prior knowledge about the situation. The assumption of complete ignorance in terms of a vague prior led us to the equivalent event of a posterior density for $\mathbf{y}_{T-d}(d)$, the (T-d)×1 vector of observations y_T, y_{T-1}, ... , and y_{d+1}, given the first d observations y_d, y_{d-1}, ... , y_1, being given by equations (3.11) and hence (3.18) of chapter three.

Hence, $$\mathbf{y}_T/\text{vague prior} \equiv \mathbf{y}_{T-d}/y_d,\ y_{d-1},\ \ldots\ ,\ y_1 = \mathbf{y}_{T-d}(d)$$

and so when σ_a^2, σ_e^2, and ϑ are also unknown,

$$\mathbf{y}_T/(\text{vague prior},\ \sigma_a^2,\sigma_e^2,\vartheta) \equiv \mathbf{y}_{T-d}(d)/(\sigma_a^2,\sigma_e^2,\vartheta)$$

whose probability density is given by:

$$\text{i.e. } p(\mathbf{y}_T/(\text{vague prior},\ \sigma_a^2,\sigma_e^2,\vartheta)) = p(\mathbf{y}_{T-d}(d)/\sigma_a^2,\sigma_e^2,\vartheta) \qquad (4.02)$$

where the distribution of $\mathbf{y}_{T-d}(d,\ \sigma_a^2,\sigma_e^2,\vartheta)$ is given from chapter three, (3.18), by,

$$\mathbf{y}_{T-d}(d,\ \sigma_a^2,\sigma_e^2,\vartheta) \sim UD\left(\ \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d;\ \mathbf{B}^{-1}.\Omega_{T-d}.(\mathbf{B}^T)^{-1}\right) \qquad (4.03)$$

from which the log-likelihood of σ_a^2, σ_e^2 and ϑ, is given by $\mathbb{LL}(\sigma_a^2,\sigma_e^2,\vartheta)$ where:

$$\mathbb{LL}(\sigma_a^2,\sigma_e^2,\vartheta) = \ln(p(\mathbf{y}_{T-d}(d,\ \sigma_a^2,\sigma_e^2,\vartheta)) \qquad (4.04)$$

where $p(\mathbf{y}_{T-d}(d,\ \sigma_a^2,\sigma_e^2,\vartheta))$ is the probability density function, introduced in (4.02).

4.21 THE ASSUMPTION OF NORMALITY

If the linearity assumptions of section **2.5** in chapter two, which lead us to (4.03), are exchanged for Normality assumptions, the undefined distribution of (4.03) becomes multivariate Normal, and hence (4.01) may be applied to (4.04) to give:

$$\mathbb{LL}(\sigma_a^2,\sigma_e^2,\vartheta) = -1/2.((T-d).\ln(2\pi) + \ln|\mathbf{B}^{-1}.\Omega_{T-d}.(\mathbf{B}^T)^{-1}|$$

$$+ (\mathbf{y}_{T-d} - \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d)^T.\mathbf{B}^T.\Omega_{T-d}^{-1}.\mathbf{B}.(\mathbf{y}_{T-d} - \mathbf{B}^{-1}.\mathbf{C}.\mathbf{y}_d)) \qquad (4.05)$$

which can be considerably simplified, because of the nature of the matrix **B**, (given by (3.04) in chapter three), to,

$$\mathbb{LL}(\sigma_a^2,\sigma_e^2,\vartheta) = -1/2.((T-d).\ln(2\pi) + \ln|\Omega_{T-d}| + \mathbf{y}_T^T.\mathbf{D}^T.\Omega_{T-d}^{-1}.\mathbf{D}.\mathbf{y}_T)$$

....(4.06)

where the matrix **D**, defined by both (3.03) and (3.04) in chapter three, and the matrix Ω_{T-d} is given by (3.10) as,

$$\Omega_{T-d} = \sigma_a^2.\mathbf{I}_{T-d} + \sigma_e^2.\mathbf{D}.\mathbf{D}^T \qquad (4.07)$$

4.3 THE CLASSICAL APPROACH

The alternative classical approach to trend estimation was described in chapter one, its general autoregressive model being defined in section **1.31**, by the vector measurement and structural equations of (1.45) and (1.46), namely,

$$\mathbf{y}_T = \mathbf{x}_T + \mathbf{e}_T, \text{ where } \mathbf{e}_T \sim UD(\varnothing;\sigma_e^2.\mathbf{I}_T) \qquad (4.08)$$

$$\mathbf{D}.\mathbf{x}_T = \mathbf{a}_{T-d}, \text{ where } \mathbf{a}_{T-d} \sim UD(\varnothing;\sigma_e^2.\mathbf{I}_{T-d}) \qquad (4.09)$$

and $\mathbb{E}[\mathbf{e}_T.\mathbf{a}_{T-d}^T] = \mathbf{\varnothing}$, where the $\mathbf{\varnothing}$'s are conformally dimensioned, vectors or matrices, of zeros.

Since the prior mean and variance of $\mathbf{x}_T$ is unspecified in this model because of the form of (4.09), the mean and variance of $\mathbf{y}_T$ is also unspecified because of (4.09). The best we can do is specify the mean and variance of $\mathbf{D}.\mathbf{y}_T$ by combining (4.08) and (4.09).

$$\text{i.e. } \mathbf{D}.\mathbf{y}_T = \mathbf{D}.\mathbf{x}_T + \mathbf{D}.\mathbf{e}_T = \mathbf{a}_{T-d} + \mathbf{D}.\mathbf{e}_T \tag{4.10}$$

From which $\mathbb{E}[\mathbf{D}.\mathbf{y}_T] = \mathbf{\varnothing}$, and $\mathbb{C}ov[\mathbf{D}.\mathbf{y}_T] = \sigma_a^2.\mathbf{I}_{T-d} + \sigma_e^2.\mathbf{D}.\mathbf{D}^T = \Omega_{T-d}$ from (4.07), hence,

$$\mathbf{D}.\mathbf{y}_T \sim UD\left(\mathbf{\varnothing};\ \Omega_{T-d} \right) \tag{4.11}$$

4.31 THE ASSUMPTION OF NORMALITY

If we now assume Normality for the distributions of (4.08) and (4.09), (4.11) becomes specified as Normal also and so the log-likelihood of σ_a^2, σ_e^2 and ϑ, $\mathbb{LL}(\sigma_a^2,\sigma_e^2,\vartheta)$, is given by (4.01), which becomes:

$$\mathbb{LL}(\sigma_a^2,\sigma_e^2,\vartheta) = \ln(p(\mathbf{D}.\mathbf{y}_T))$$

$$= -1/2.((T-d).\ln(2\pi) + \ln|\Omega_{T-d}| + \mathbf{y}_T^T.\mathbf{D}^T.\Omega_{T-d}^{-1}.\mathbf{D}.\mathbf{y}_T) \tag{4.12}$$

which is exactly the same equation as that produced using the State Space approach, i.e. (4.06).

4.4 MAXIMISATION OF THE LIKELIHOOD FUNCTION

We shall find it useful, in maximising (4.06) or (4.12), to replace σ_a^2 by $\omega.\sigma_e^2$, where the variance ratio, $\omega = \sigma_a^2/\sigma_e^2$, was first defined in

equation (1.36) of chapter one. Since the only occurrence of σ_a^2 is in the matrix $\Omega_{T-d} = \sigma_a^2.\mathbf{I}_{T-d} + \sigma_e^2.\mathbf{D}.\mathbf{D}^T$, this only requires replacing Ω_{T-d} by $\sigma_e^2.\Omega_\omega$ where the matrix Ω_ω is defined by (4.13) as:

$$\Omega_\omega = \omega.\mathbf{I}_{T-d} + \mathbf{D}.\mathbf{D}^T \tag{4.13}$$

The log-likelihood function becomes $\mathbb{LL}(\omega,\sigma_e^2,\vartheta)$, where:

$$\mathbb{LL}(\omega,\sigma_e^2,\vartheta) = -1/2.\Big((T-d).\ln(2\pi) + (T-d).\ln(\sigma_e^2) + \log|\Omega_\omega| + \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T/\sigma_e^2 \Big) \tag{4.14}$$

The final stage is, in theory, quite straightforward, namely to differentiate (4.14) with respect to ω and σ_e^2, set the result to zero, and then solve the resulting equations to give the required log-likelihood estimates of ω and σ_e^2.

Differentiation of (4.14) w.r.t σ_e^2 is straightforward and leads to the relationship, (where of course σ_e^2 is now an M.L. estimate),

$$\sigma_e^2 = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T/(T-d) \tag{4.15}$$

Differentiation of (4.14) w.r.t ω is not quite so obvious but still straightforward by utilising the matrix differentiation relationships,

$$\partial\ln|\Omega_\omega|/\partial\omega = \mathbb{TR}[\Omega_\omega^{-1}.\partial\Omega_\omega/\partial\omega],$$

and $$\partial(\mathbf{y}_T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T)/\partial\omega = -\mathbf{y}_T.\mathbf{D}^T.\Omega_\omega^{-1}.\partial\Omega_\omega/\partial\omega.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T$$

where $\mathbb{TR}[M]$ stands for the trace of M, and leads to the relationship.

$$\mathbb{TR}[\Omega_\omega^{-1}] = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T/\sigma_e^2 \tag{4.16}$$

which on utilising (4.15), removes the parameter σ_e^2, to give,

$$(T-d).\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T = \mathrm{TR}[\Omega_\omega^{-1}].\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T \qquad (4.17)$$

Solving (4.17), (see next chapter), gives us a value for ω, which when substituted in (4.15) gives us σ_e^2, and hence $\sigma_a^2 = \omega.\sigma_e^2$. If (4.17) gives more than one solution, we choose that which maximises the log-likelihood, i.e. (4.14).

4.5 A NON-LIKELIHOOD APPROACH USING THE QUADRATIC FORM

Following the best traditions of lateral thinking, we now explore another avenue which might give us further insights into the problem.

4.51 VARIANCE ESTIMATION USING THE QUADRATIC FORM

The simplest form of statistical estimator for a parameter having the same dimension as the observations, is a *linear* function of the observations. When we come to estimate parameters which are essentially squared measures in this respect, the simplest form of estimator is a *quadratic* function of the observations. The estimation of a *variance* is in this category.

The general quadratic function of a set of observations y_1, y_2, ... , y_T, can be written as $Q(y_1,y_2,\ldots,y_T)$, which must also be the simplest form of a general variance estimator, $\hat{\sigma}^2$, i.e.

$$\hat{\sigma}^2 = Q(y_1,y_2,\ldots,y_T) = \sum_{i=1}^{T}\sum_{j=1}^{T} c_{ij}.y_i.y_j \qquad (4.18)$$

where the coefficients c_{ij} can be arbitrarily chosen. In matrix terms, this can be written,

$$\hat{\sigma}^2 = \mathbf{y}_T^T.\mathbf{C}.\mathbf{y}_T \qquad (4.19)$$

where $\mathbf{y}_T$ is the vector of observations i.e. $\mathbf{y}_T = (y_1, y_2, \ldots, y_T)^T$, and $\mathbf{C}$ is a square matrix whose ij th element is c_{ij}.

The matrix $\mathbf{C}$ need not be symmetrical but can always be chosen to be so, since from (4.18) the coefficient of the term $y_i.y_j$ is $(c_{ij} + c_{ji})$, implying that c_{ij} can always be chosen equal to c_{ji} without affecting the result.

We have already seen, earlier in this chapter, that there is no unconstrained relationship between the data vector, $\mathbf{y}_T$ and the stochastic parts of the model, $\mathbf{e}_T$ and $\mathbf{a}_{T-d}$. The only direct way we have of relating the two being (4.10), i.e. via the random vector $\mathbf{D}.\mathbf{y}_T$ defined in (4.11).

The implication this has for the estimation of model variance parameters is that the estimator in (4.19) needs to have a general form given by:

$$\hat{\sigma}^2 = \mathbf{y}_T^T.\mathbf{D}^T.\mathbf{M}.\mathbf{D}.\mathbf{y}_T \tag{4.20}$$

where $\mathbf{C} = \mathbf{D}^T.\mathbf{M}.\mathbf{D}$ and $\mathbf{M}$ is again an arbitrary square matrix of dimension T-d whose ij th element is m_{ij}.

Suppose we now impose the condition that the undefined distribution (4.11) is multivariate Normal, (implying that all residuals were also multivariate Normal).

Then, from appendix B, $\hat{\sigma}^2$ would have mean, $\mathbb{E}[\hat{\sigma}^2]$, and variance, $\mathbb{V}[\hat{\sigma}^2]$, given by:

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{TR}\left[\, \mathbf{M}.\Omega_{T-d} \,\right] \tag{4.21}$$

and,

$$\mathbb{V}[\hat{\sigma}^2] = 2.\mathbb{TR}\left[\, \mathbf{M}.\Omega_{T-d}.\mathbf{M}.\Omega_{T-d} \,\right] \tag{4.22}$$

4.52 MINIMUM VARIANCE CONDITIONALLY UNBIASED ESTIMATION

To obtain a minimum variance, unbiased estimator, $\hat{\sigma}^2$, of a variance σ^2, we need to minimise the variance in (4.22) with respect to each of the elements, m_{ij}, of $\mathbf{M}$, subject to the condition:

$$\mathbb{TR}\left[\mathbf{M}.\Omega_{T-d} \right] = \sigma^2 \tag{4.23}$$

The constraint in (4.23) can be incorporated into the minimisation of the variance in (4.22) using the Lagrangian multiplier, $4.\lambda$. Hence, the function of all elements m_{ij} of $\mathbf{M}$, which we need to minimise, with respect to each m_{ij} is given by $f(m_{ij};\ i,j = 1 \text{ to } T-d)$, or more concisely $f(\mathbf{M})$, where:

$$f(\mathbf{M}) = 2.\mathbb{TR}\left[\mathbf{M}.\Omega_{T-d}.\mathbf{M}.\Omega_{T-d} \right] - 4.\lambda.\left(\mathbb{TR}\left[\mathbf{M}.\Omega_{T-d} \right] - \sigma^2 \right) \tag{4.24}$$

From appendix C, (C6) and (C7), we have the following two results,

$$\partial/\partial\mathbf{M} \left\{ \mathbb{TR}\left[\mathbf{M}.\Omega_{T-d} \right] \right\} = \Omega_{T-d} \tag{4.25}$$

$$\partial/\partial\mathbf{M} \left\{ \mathbb{TR}\left[\mathbf{M}.\Omega_{T-d}.\mathbf{M}.\Omega_{T-d} \right] \right\} = 2.\Omega_{T-d}.\mathbf{M}^T.\Omega_{T-d} \tag{4.26}$$

Therefore differentiating (4.24), using (4.25) and (4.26), we have

$$\partial/\partial\mathbf{M} \left\{ f(\mathbf{M}) \right\} = 4.\Omega_{T-d}.\mathbf{M}^T.\Omega_{T-d} - 4.\lambda.\Omega_{T-d} \tag{4.27}$$

Hence at the optimum, realising $\mathbf{M}$ is symmetric, (4.27) gives,

$$\mathbf{M} = \mathbf{M}^T = \lambda.\Omega_{T-d}^{-1} \tag{4.28}$$

Using (4.23) and (4.28), gives a value for λ, i.e.

$$\mathbb{TR}\left[\mathbf{M}.\Omega_{T-d} \right] = \lambda.\mathbb{TR}[\mathbf{I}_{T-d}] = \lambda.(T-d) = \sigma^2 \qquad (4.29)$$

which can be, using (4.13), ($\Omega_{T-d} = \sigma_e^2.\Omega_\omega$), substituted into (4.28) to complete the solution, i.e.

$$\mathbf{M} = \sigma^2.\Omega_{T-d}^{-1}/(T-d) = (\sigma^2/\sigma_e^2).\Omega_\omega^{-1}/(T-d) \qquad (4.30)$$

Utilisation of (4.20), with $\sigma^2 = \sigma_e^2$, gives us a minimum variance unbiased estimator for σ_e^2, i.e.

$$\hat{\sigma}_e^2 = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T/(T-d) \qquad (4.31)$$

which is the same as the one given by the likelihood function in (4.15),

Similarly for $\sigma^2 = \sigma_a^2$, we get,

$$\hat{\sigma}_a^2 = \omega.\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T/(T-d) = \omega.\hat{\sigma}_e^2 \qquad (4.32)$$

Both estimates in (4.31) and (4.32) are functions of the variance ratio, ω, and hence require its pre-specification. We therefore refer to them as minimum variance conditionally unbiased estimates.

4.53 MINIMUM VARIANCE UNCONDITIONALLY UNBIASED ESTIMATION

To obtain a minimum variance, unconditionally unbiased estimator, $\hat{\sigma}^2$, of the variance σ^2, where,

$$\sigma^2 = \alpha.\sigma_a^2 + \beta.\sigma_e^2 \qquad (4.33)$$

we take advantage of the fact that (4.21) can be written,

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{TR}\left[\mathbf{M}.\Omega_{T-d} \right] = \sigma_a^2.\mathbb{TR}\left[\mathbf{M} \right] + \sigma_e^2.\mathbb{TR}\left[\mathbf{M}.\mathbf{D}.\mathbf{D}^T \right] \qquad (4.34)$$

Again we need to minimise the variance in (4.22) with respect to each of the elements, m_{ij}, of $\mathbf{M}$, but now subject to the two conditions:

$$\mathbb{TR}\left[\mathbf{M} \right] = \alpha \qquad \text{and} \qquad \mathbb{TR}\left[\mathbf{M}.\mathbf{D}.\mathbf{D}^T \right] = \beta \qquad (4.35)$$

The proof now requires two lagrangian multipliers $4.\lambda$ and $4.\mu$ to incorporate the constraints of (4.35). Hence, the function of all elements m_{ij} of $\mathbf{M}$, which we need to minimise, with respect to each m_{ij} is given by $f(\mathbf{M})$, where:

$$f(\mathbf{M}) = 2.\mathbb{TR}\left[\mathbf{M}.\Omega_{T-d}.\mathbf{M}.\Omega_{T-d}\right] - 4.\lambda.\left(\mathbb{TR}\left[\mathbf{M}\right] - \alpha\right) - 4.\mu.\left(\mathbb{TR}\left[\mathbf{M}.\mathbf{D}.\mathbf{D}^T\right] - \beta\right) \qquad ....(4.36)$$

From appendix C, (C4), (C6) and (C8) we have the following three results,

$$\partial/\partial\mathbf{M}\left\{ \mathbb{TR}\left[\mathbf{M} \right] \right\} = \mathbf{I}_{T-d} \qquad (4.37)$$

$$\partial/\partial\mathbf{M}\left\{ \mathbb{TR}\left[\mathbf{M}.\mathbf{D}.\mathbf{D}^T \right] \right\} = \mathbf{D}.\mathbf{D}^T \qquad (4.38)$$

$$\partial/\partial\mathbf{M}\left\{ \mathbb{TR}\left[\mathbf{M}.\Omega_{T-d}.\mathbf{M}.\Omega_{T-d} \right] \right\} = 2.\Omega_{T-d}.\mathbf{M}^T.\Omega_{T-d} \qquad (4.39)$$

Therefore differentiating (4.36), using (4.37) to (4.39), we have

$$\partial/\partial \mathbf{M} \left\{ f(\mathbf{M}) \right\} = 4.\Omega_{T-d}.\mathbf{M}^T.\Omega_{T-d} - 4.\lambda.\mathbf{I}_{T-d} - 4.\mu.\mathbf{D}.\mathbf{D}^T \qquad (4.40)$$

Hence at the optimum, (4.40) gives (4.41) and then finally (4.42), where

$$\Omega_{T-d}.\mathbf{M}^T.\Omega_{T-d} = \lambda.\mathbf{I}_{T-d} + \mu.\mathbf{D}.\mathbf{D}^T \qquad (4.41)$$

$$\mathbf{M} = \mathbf{M}^T = \lambda.\Omega_{T-d}^{-2} + \mu.\Omega_{T-d}^{-1}.\mathbf{D}.\mathbf{D}^T.\Omega_{T-d}^{-1} \qquad (4.42)$$

As in section **4.4**, we shall find it more convenient to work with the matrix Ω_ω, defined in (4.13), where $\Omega_{T-d} = \sigma_e^2.\Omega_\omega$. Hence (4.42) becomes:

$$\mathbf{M} = \lambda^*.\Omega_\omega^{-2} + \mu^*.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{D}^T.\Omega_\omega^{-1} \qquad (4.43)$$

$$\text{where } \lambda^* = \lambda/\sigma_e^4 \text{ and } \mu^* = \mu/\sigma_e^4 \qquad (4.44)$$

From (4.13) we have,

$$\mathbf{D}.\mathbf{D}^T = (\Omega_\omega - \omega.\mathbf{I}_{T-d}) \qquad (4.45)$$

Hence from (4.43) and (4.45), we have,

$$\mathbf{M} = \lambda^*.\Omega_\omega^{-2} + \mu^*.\left(\Omega_\omega^{-1} - \omega.\Omega_\omega^{-2} \right) \qquad (4.46)$$

Hence,

$$\mathbb{TR}\left[\mathbf{M} \right] = \lambda^*.\mathbb{TR}\left[\Omega_\omega^{-2} \right] + \mu^*.\mathbb{TR}\left[\Omega_\omega^{-1} - \omega.\Omega_\omega^{-2} \right] \qquad (4.47)$$

and, again using (4.45),

$$\mathbb{TR}\left[\mathbf{M}.\mathbf{D}.\mathbf{D}^{T} \right] = \mathbb{TR}\left[\mathbf{M}.(\Omega_{\omega} - \omega.\mathbf{I}_{T\text{-}d}) \right] = \mathbb{TR}\left[\mathbf{M}.\Omega_{\omega} \right] - \omega.\mathbb{TR}\left[\mathbf{M} \right] \quad (4.48)$$

Simplifying (4.48), using (4.46), we get:

$$\mathbb{TR}\left[\mathbf{M}.\mathbf{D}.\mathbf{D}^{T} \right] = \lambda^{*}.\mathbb{TR}\left[\Omega_{\omega}^{-1} - \omega.\Omega_{\omega}^{-2} \right] + \mu^{*}.\mathbb{TR}\left[\mathbf{I}_{T\text{-}d} - 2.\omega.\Omega_{\omega}^{-1} + \omega^{2}.\Omega_{\omega}^{-2} \right]$$

....(4.49)

Writing $\mathbb{TR}[\Omega_{\omega}^{-k}]$ as $\mathbb{T}_{k}$, for k=0 to 2, in (4.47) and (4.49), equation (4.35) can be written,

$$\begin{Bmatrix} \alpha \\ \beta \end{Bmatrix} = \begin{Bmatrix} \mathbb{T}_{2} & \mathbb{T}_{1}-\omega.\mathbb{T}_{2} \\ \mathbb{T}_{1}-\omega.\mathbb{T}_{2} & \mathbb{T}_{0}-2.\omega.\mathbb{T}_{1}+\omega^{2}.\mathbb{T}_{2} \end{Bmatrix}.\begin{Bmatrix} \lambda^{*} \\ \mu^{*} \end{Bmatrix} \quad (4.50)$$

Inverting (4.50) we get,

$$\begin{Bmatrix} \lambda^{*} \\ \mu^{*} \end{Bmatrix} = \begin{Bmatrix} \mathbb{T}_{0}-2.\omega.\mathbb{T}_{1}+\omega^{2}.\mathbb{T}_{2} & -(\mathbb{T}_{1}-\omega.\mathbb{T}_{2}) \\ -(\mathbb{T}_{1}-\omega.\mathbb{T}_{2}) & \mathbb{T}_{2} \end{Bmatrix}.\begin{Bmatrix} \alpha \\ \beta \end{Bmatrix} \Big/ \left(\mathbb{T}_{0}.\mathbb{T}_{2}-\mathbb{T}_{1}^{2} \right) \quad (4.51)$$

Finally, substituting (4.51) into (4.46), the expression for **M** becomes:

$$\mathbf{M} = \left(\Omega_{\omega}^{-2},\ \Omega_{\omega}^{-1}-\omega.\Omega_{\omega}^{-2}\right)\begin{Bmatrix} \mathbb{T}_{0}-2.\omega.\mathbb{T}_{1}+\omega^{2}.\mathbb{T}_{2} & -(\mathbb{T}_{1}-\omega.\mathbb{T}_{2}) \\ -(\mathbb{T}_{1}-\omega.\mathbb{T}_{2}) & \mathbb{T}_{2} \end{Bmatrix}.\begin{Bmatrix} \alpha \\ \beta \end{Bmatrix} \Big/ \left(\mathbb{T}_{0}.\mathbb{T}_{2}-\mathbb{T}_{1}^{2} \right)$$

....(4.52)

Since $\mathbf{M}$ is a function of ω in (4.52), it would appear that the estimator is again conditionally unbiased. However if we choose $\omega = \nu \neq \sigma_a^2/\sigma_e^2$, such that Ω_ω becomes Ω_ν etc. in (4.52), it is a straightforward matter of substitution to show that,

$$\mathbb{E}\left[\hat{\sigma}^2 \right] = \mathbb{TR}\left[\mathbf{M}(\omega=\nu).\Omega_{T-d} \right] = \left[\mathbf{M}(\omega=\nu).(\sigma_a^2.\mathbf{I}_{T-d} + \sigma_e^2.\mathbf{D}.\mathbf{D}^T) \right]$$

$$= \mathbb{TR}\left[\mathbf{M}(\omega=\nu).(\sigma_a^2.\mathbf{I}_{T-d} - \sigma_e^2.(\Omega_\nu - \nu.\mathbf{I}_{T-d}) \right] = \alpha.\sigma_a^2 + \beta.\sigma_e^2 = \sigma^2 \qquad (4.53)$$

Hence $\hat{\sigma}^2$ is <u>unconditionally</u> unbiased.

4.531 ESTIMATION OF THE MEASUREMENT VARIANCE

Using (4.20), an unbiased estimator of σ_e^2 with minimum variance is given by $\hat{\sigma}_e^2$, where:

$$\hat{\sigma}_e^2 = \mathbf{y}_T^T.\mathbf{D}^T.\mathbf{M}_e.\mathbf{D}.\mathbf{y}_T \qquad (4.54)$$

and the matrix $\mathbf{M}_e$ is given by setting $\alpha = 0$, and $\beta = 1$ in (4.52), since $\sigma^2 = \sigma_e^2$ for these values in (4.33). Therefore,

$$\mathbf{M}_e = \left\{ \mathbb{T}_2.\Omega_\omega^{-1} - \mathbb{T}_1.\Omega_\omega^{-2} \right\} / \left(\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1^2 \right) \qquad (4.55)$$

4.532 ESTIMATION OF THE STRUCTURAL VARIANCE

Again using (4.20) an unbiased estimator of σ_a^2 with minimum variance is given by $\hat{\sigma}_a^2$, where:

$$\hat{\sigma}_a^2 = \mathbf{y}_T^T.\mathbf{D}^T.\mathbf{M}_a.\mathbf{D}.\mathbf{y}_T \qquad (4.56)$$

and, because of (4.33), the matrix $\mathbf{M}_a$ is given by setting $\alpha = 1$, and $\beta = 0$ in (4.52), i.e.

$$\mathbf{M}_a = \left\{ \omega.(\mathbb{T}_2.\Omega_\omega^{-1} - \mathbb{T}_1.\Omega_\omega^{-2}) + \mathbb{T}_0.\Omega_\omega^{-2} - \mathbb{T}_1.\Omega_\omega^{-1} \right\} / \left(\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1^2 \right) \quad (4.57)$$

4.533 ESTIMATION OF THE VARIANCE RATIO

Substituting (4.55) into (4.57) gives,

$$\mathbf{M}_a = \omega.\mathbf{M}_e + (\mathbb{T}_0.\Omega_\omega^{-2} - \mathbb{T}_1.\Omega_\omega^{-1})/(\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1^2) \quad (4.58)$$

which after applying (4.54) and (4.56) gives,

$$\hat{\sigma}_a^2 = \omega.\hat{\sigma}_e^2 + (\mathbb{T}_0.\mathbb{Q}_2 - \mathbb{T}_1.\mathbb{Q}_1)/(\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1^2) \quad (4.59)$$

$$\text{where } \mathbb{Q}_k = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-k}.\mathbf{D}.\mathbf{y}_T \quad (4.60)$$

Also applying (4.54) to (4.55), we get,

$$\hat{\sigma}_e^2 = (\mathbb{T}_2.\mathbb{Q}_1 - \mathbb{T}_1.\mathbb{Q}_2)/(\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1^2) \quad (4.61)$$

Hence combining (4.59) and (4.61) gives,

$$\hat{\omega} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_e^2} = \omega + \frac{\mathbb{T}_0.\mathbb{Q}_2 - \mathbb{T}_1.\mathbb{Q}_1}{\mathbb{T}_2.\mathbb{Q}_1 - \mathbb{T}_1.\mathbb{Q}_2} \quad (4.62)$$

For (4.62) to be true for $\omega = \hat{\omega}$, this requires $\mathbb{T}_0.\mathbb{Q}_2 = \mathbb{T}_1.\mathbb{Q}_1$, i.e.

$$(T-d).\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T = \mathbb{TR}[\Omega_\omega^{-1}].\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T \quad (4.63)$$

Hence by solving (4.63) we will obtain the required estimate for ω, $\hat{\omega}$, which can then be substituted into (4.61) to give $\hat{\sigma}_e^2$ and therefore $\hat{\sigma}_a^2 = \hat{\omega}.\hat{\sigma}_e^2$.

Note that condition (4.63) is exactly the same as that produced from Maximum Likelihood in (4.17).

Also substituting (4.63) into (4.61), we obtain,

$$\hat{\sigma}_e^2 = \mathbb{Q}_1/\mathbb{T}_0 = \mathbb{Q}_2/\mathbb{T}_1 \qquad (4.64)$$

which again is exactly the same result as obtained using Maximum Likelihood in (4.15).

Hence the above, Minimum Variance estimates are identical to those produced using Maximum Likelihood.

Note that equation (4.62) can be thought of as a recurrence relation for estimating ω, i.e.

$$\omega_{OUT} = \omega_{IN} + \delta(\omega_{IN}) \qquad (4.65)$$

The iteration of (4.65) is exploited in the next chapter.

4.54 THE ASSUMPTION OF NORMALITY

For this section, **4.5**, it was apparently necessary to include the condition that the distribution in (4.11) was multivariate Normal. In fact the only result of the section which did require this result was (4.22), which, upon inspection of its derivation in (*Searle,1971,chapter 2*) cited in appendix B, only requires the lesser condition that the unspecified third and forth moments of the distribution in (4.11) have the same relationship to the first two as those of a multivariate Normal distribution.

The third moment, being zero, essentially imposes the condition that the distribution in (4.11) is symmetric, which still leaves some room for non-Normal variations in (4.11); it is, unfortunately, the imposition of the forth moment condition which essentially limits us to distributions which are so close to Normality that it serves little practical purpose to define them as otherwise.

4.6 WHITTAKER'S SUM OF SQUARES FUNCTION

From equation (1.10) of chapter one, Whittaker's, (weighted sum of squares), function ψ, is defined as:

$$\psi = (\mathbf{y}_T-\mathbf{x}_T)^T.(\mathbf{y}_T-\mathbf{x}_T) + 1/\omega.\mathbf{x}_T^T.\mathbf{D}^T.\mathbf{D}.\mathbf{x}_T \qquad (4.66)$$

The optimal value of $\mathbf{x}_T$, $\hat{\mathbf{x}}_T$, which minimises ψ, (for a given ω), is given by (1.14) of chapter one as:

$$\hat{\mathbf{x}}_T = (\mathbf{I}_T+1/\omega.\mathbf{D}^T.\mathbf{D})^{-1}\mathbf{y}_T \qquad (4.67)$$

Substitution of (4.67) into (4.66) is performed using steps (4.68) to (4.74), below. Rewriting (4.67), we have,

$$(\mathbf{I}_T+1/\omega.\mathbf{D}^T.\mathbf{D}).\hat{\mathbf{x}}_T = \mathbf{y}_T \qquad (4.68)$$

and hence,

$$\mathbf{y}_T-\hat{\mathbf{x}}_T = 1/\omega.\mathbf{D}^T.\mathbf{D}.\hat{\mathbf{x}}_T \qquad (4.69)$$

From (4.69) it follows that:

$$(\mathbf{y}_T-\hat{\mathbf{x}}_T)^T.(\mathbf{y}_T-\hat{\mathbf{x}}_T) = 1/\omega.\mathbf{y}_T^T.\mathbf{D}^T.\mathbf{D}.\hat{\mathbf{x}}_T - 1/\omega.\hat{\mathbf{x}}_T^T.\mathbf{D}^T.\mathbf{D}.\hat{\mathbf{x}}_T \qquad (4.70)$$

which on substitution into (4.66) gives,

$$\psi(\mathbf{x}_T=\hat{\mathbf{x}}_T) = \hat{\psi} = 1/\omega.\mathbf{y}_T^T.\mathbf{D}^T.\mathbf{D}.\hat{\mathbf{x}}_T \tag{4.71}$$

Furthermore pre-multiplying (4.68) by $\mathbf{D}$, (when $\mathbf{D}$ is in its general form of (1.44), chapter one, having dimensions (T-d)×T, we get,

$$\mathbf{D}.\mathbf{y}_T = 1/\omega.(\omega.\mathbf{I}_{T-d} + \mathbf{D}.\mathbf{D}^T).\mathbf{D}.\hat{\mathbf{x}}_T = 1/\omega.\Omega_\omega.\mathbf{D}.\hat{\mathbf{x}}_T \tag{4.72}$$

where Ω_ω was defined in (4.13).

From which it follows that,

$$1/\omega.\mathbf{D}.\hat{\mathbf{x}}_T = \Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T \tag{4.73}$$

which finally, on substitution into (4.71), remembering (4.60), gives,

$$\psi(\mathbf{x}_T=\hat{\mathbf{x}}_T) = \hat{\psi} = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T = \mathbb{Q}_1(\omega) \tag{4.74}$$

Notice that the form of the function $\hat{\psi}$ in (4.74), i.e. $\mathbb{Q}_1(\cdot)$ defined in (4.64), has appeared quite a lot in this chapter. It occurs in the likelihood function (4.14), in estimates for σ_e^2, and therefore σ_a^2, (4.15), (4.31) and (4.32) and in the equation for estimating ω, (4.17) and (4.63). As we saw at the beginning of chapter one, ψ is a weighted sums of squares function, weighting together the sum of squared errors/residuals associated with fidelity or measurement, (SSe), and the sum of squared errors/residuals associated with smoothness, i.e. the structural errors, (SSa). Hence we could write, Whittaker's function, ψ, using (4.66) as,

$$\psi = SSe + 1/\omega.SSa \tag{4.75}$$

where,

$$SSe = (\mathbf{y}_T - \mathbf{x}_T)^T . (\mathbf{y}_T - \mathbf{x}_T) \tag{4.76}$$

$$SSa = \mathbf{x}_T^T . \mathbf{D}^T . \mathbf{D} . \mathbf{x}_T \tag{4.77}$$

And hence, from (4.74),

$$\hat{\psi} = \psi(\mathbf{x}_T = \hat{\mathbf{x}}_T) = \mathbf{y}_T^T . \mathbf{D}^T . \Omega_\omega^{-1} . \mathbf{D} . \mathbf{y}_T = \hat{SSe} + 1/\omega . \hat{SSa} = = \mathbb{Q}_1(\omega)$$

$$\text{where } \hat{SSe} = SSe(\mathbf{x}_T = \hat{\mathbf{x}}_T), \text{ and } \hat{SSa} = SSa(\mathbf{x}_T = \hat{\mathbf{x}}_T) \tag{4.78}$$

Using (4.73) and (4.77), we have,

$$\hat{SSa} = SSa(\mathbf{x}_T = \hat{\mathbf{x}}_T) = \hat{\mathbf{x}}_T^T . \mathbf{D}^T . \mathbf{D} . \hat{\mathbf{x}}_T = \omega^2 . \mathbf{y}_T^T . \mathbf{D}^T . \Omega_\omega^{-2} . \mathbf{D} . \mathbf{y}_T = \omega^2 . \mathbb{Q}_2(\omega) \tag{4.79}$$

The form of the function $\mathbf{y}_T^T . \mathbf{D}^T . \Omega_\omega^{-2} . \mathbf{D} . \mathbf{y}_T$ in (4.79), i.e. $\mathbb{Q}_2(\cdot)$ defined in (4.60), also appears in this chapter, in (4.17) and (4.63).

Substituting (4.78) and (4.79) into (4.63) or (4.64), the equation for estimating ω, gives us,

$$(T-d) . 1/\omega . \hat{SSa} = \omega . \mathbb{TR}[\Omega_\omega^{-1}] . (\hat{SSe} + 1/\omega . \hat{SSa}) \tag{4.80}$$

which gives us an alternative approach, (via the estimation of $\hat{\mathbf{x}}_T$), to estimating ω.

4.61 INTUITIVE ESTIMATION

Writing (4.76) and (4.77), (when $\mathbf{x}_T = \hat{\mathbf{x}}_T$), as summations of the individual vector elements, we have,

$$\hat{SSe} = \sum_{t=1}^{t=T} (y_t - \hat{x}_t)^2 = \sum_{t=1}^{t=T} \hat{e}_t^{\,2} \tag{4.81}$$

$$\hat{SSa} = \sum_{t=1}^{t=T-d} (\hat{x}_t - \vartheta_1.\hat{x}_{t-1} - \vartheta_2.\hat{x}_{t-2} - \ldots . \vartheta_d.\hat{x}_{t-d})^2 = \sum_{t=1}^{t=T-d} \hat{a}_t^2 \qquad (4.82)$$

Intuitively, therefore, we might consider estimating σ_e^2 by $\hat{SSe}/T$, σ_a^2 by $\hat{SSa}/(T-d)$ and hence ω, $(= \sigma_a^2/\sigma_e^2)$ by the ratio of these two estimates i.e. $T.\hat{SSa}/(T-d)/\hat{SSe}$.

$\hat{SSa}$ is given by (4.79). Hence an intuitive estimate for σ_a^2 is given by:

$$\hat{SSa}/(T-d) = \omega^2.Q_2(\omega)/(T-d) = \omega^2.\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T/(T-d) \qquad (4.83)$$

$\hat{SSe}$ is given by substituting (4.79) into (4.78); thus,

$$\hat{SSe} = Q_1(\omega) - \omega.Q_2(\omega) = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T - \omega.\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T \qquad (4.84)$$

An intuitive estimate for σ_e^2 is therefore given by,

$$\hat{SSe}/T = (Q_1(\omega) - \omega.Q_2(\omega))/T = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T/T - \omega.\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T/T$$

$$\ldots . (4.85)$$

Since $\hat{SSe}/T$ and $\hat{SSa}/(T-d)/\omega$ are two estimates for σ_e^2, we can equate these to obtain an "intrinsic" equivalent equation to (4.17) or (4.63) for estimating ω, i.e. using (4.83) and (4.85),

$$\omega.Q_2(\omega)/(T-d) = (Q_1(\omega) - \omega.Q_2(\omega))/T \qquad (4.86)$$

i.e.

$$\omega.(2T-d).Q_2(\omega) = (T-d).Q_1(\omega) \qquad (4.87)$$

i.e.

$$\omega.(2T-d).\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-2}.\mathbf{D}.\mathbf{y}_T = (T-d).\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T \qquad (4.88)$$

As we can see equations (4.83), (4.85) and (4.88) differ quite significantly from previous "optimal" results and lead to quite different estimates. The extent of the differences were confirmed by some initial trials with simulated, (known variance), data, where their performance, as opposed to the "optimal estimates" was positively misleading, and, as such, they were not pursued further.

However, it is cautionary to note that, unlike many other estimation procedures in which theory simply confirms intuition, (or at least only slightly amends it), this is a situation in which intuition would not seem to be at all reliable, although why this should be the case is not particularly obvious.

4.7 SUMMARY OF RESULTS

In this chapter we investigated the estimation of the residual variances σ_a^2 and σ_e^2, firstly using Maximum Likelihood, where we showed that both the Classical approach and the State Space approach, based on the assumption of a vague prior), led to the same likelihood function, and then by considering a Minimum Variance approach which utilised a quadratic form for the estimators. In both instances these were shown to give identical estimates. The chapter ended with a cautionary note on using intuitive, but non-optimal, estimates.

In the next chapter we shall be combining the best features of both of the two optimal estimation procedures looked at in this chapter to produce an algorithm which will efficiently implement their estimation in practice.

ESTIMATION OF RESIDUAL VARIANCES II

In many situations we wish to investigate the appropriateness of trend models which are <u>invariant</u>, (see chapter one section 1.3), to simple trends e.g. constant (d=1), linear (d=2), quadratic (d=3) etc. where the parameter d is the autoregressive lag.

For these cases the individual values of the autoregressive parameters i.e. ϑ_1, ϑ_2, ... , ϑ_d, (the latter being denoted by the set or vector $\underline{\vartheta}_d$, or simply ϑ_d or ϑ if its meaning is clear), are given by the coefficients of x^r in the expansion of $-(1-x)^d$, i.e.

$$\vartheta_r = (-1)^{r+1}.d!/(r!(d-r)!) \qquad r=1, 2, \ldots , d \qquad (5.01)$$

Alternatively it may simply be the case that we wish to inestigate the effects of a particular model of chosen autoregressive parameters. In either event we can regard the parameters d and $\underline{\vartheta}_d$ as pre-specified.

This chapter build on the results of chapter four to implement a procedure for estimating the remaining unknown parameters of such models, namely the residual variances, σ_e^2 and σ_a^2. Two scenarios are considered, (i) when the variance ratio, $\omega = \sigma_a^2/\sigma_e^2$, is known and (ii) when the variance ratio is unknown.

5.1 SCENARIO 1: ESTIMATION OF RESIDUAL VARIANCES GIVEN ω and $\underline{\vartheta}_d$

In chapter one we saw that, whereas the calculation of the mean squared error of the trend values, (e.g. equation (1.38)), requires estimates for both residual variances, the calculation of the trend values themselves, (e.g. equation (1.37)), needs only their ratio, ω.

A modeller may therefore inspect the behaviour of different trends prior to carrying out a full analysis, (as in figure 1.1 of chapter

one for example), and from these choose the one which his qualitative prior knowledge of the situation deems to be most appropriate.

In practice this would mean examining various models on the basis of their smoothness and hence choosing one by the specification of its smoothness parameter "p", which, since $\omega=(1-p)/p$, (equation 1.04, chapter one), is equivalent to the Bayesian practice of invoking an "a priori" belief on the value of ω.

Given this scenario the only parameters, (apart from the trend values themselves of course), which require estimation are the residual variances, σ_e^2 and σ_a^2, which are automatically given by the (Maximum Likelihood) conditionally unbiased estimators of equations 4.31 and 4.32 of chapter four, namely,

$$\hat{\sigma}_e^2 = \mathbb{Q}_1(\omega)/\mathbb{T}_0 \tag{5.02}$$

$$\hat{\sigma}_a^2 = \omega.\mathbb{Q}_1(\omega)/\mathbb{T}_0 \tag{5.03}$$

$$\text{where } \mathbb{Q}_1(\omega) = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}_T \text{ and } \mathbb{T}_0 = T-d \tag{5.04}$$

Note also that equation 4.24 of chapter four would also produce an appropriate estimate of the variances of these parameters.

5.2 SCENARIO 2: ESTIMATION OF RESIDUAL VARIANCES GIVEN $\underset{\sim}{\vartheta}_d$

In this scenario the autoregressive parameters contained in $\underset{\sim}{\vartheta}_d$ are the only parameters to be specified, a priori, which in practice requires the estimation of ω in addition to the residual variances, σ_e^2 and σ_a^2.

5.21 THE ESTIMATION OF ω

Chapter four suggests two approaches to the estimation of ω given $\underset{\sim}{\vartheta}_d$, both of which are shown to lead to the same problem i.e. to find the

value(s) of ω which satisfy the following condition,

$$\mathbb{Q}_2(\omega)/\mathbb{Q}_1(\omega) = \mathbb{T}_1(\omega)/\mathbb{T}_0 \tag{5.05}$$

$$\text{where } \mathbb{Q}_k(\omega) = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-k}.\mathbf{D}.\mathbf{y}_T \tag{5.06}$$

$$\text{and } \mathbb{T}_k(\omega) = \mathbb{TR}[\Omega_\omega^{-k}] \tag{5.07}$$

The first approach is to maximise the log-likelihood function, (equation 4.14), with respect to ω and concentrate the resulting derivative by substituting the Maximum Likelihood estimate of σ_e^2, (equation 4.15), for σ_e^2. The resulting concentrated derivative $\partial\mathbb{LL}/\partial\omega$ is given by,

$$\partial\mathbb{LL}/\partial\omega = \mathbb{T}_0/2.(\mathbb{Q}_2(\omega)/\mathbb{Q}_1(\omega) - \mathbb{T}_1(\omega)/\mathbb{T}_0) \tag{5.08}$$

which when $\partial\mathbb{LL}/\partial\omega=0$ gives (5.05). This suggests a solution procedure based on Newton's formula but utilising first and second function derivatives, namely

$$\omega_{OUT} = \omega_{IN} - \frac{\partial\mathbb{LL}/\partial\omega(\omega=\omega_{IN})}{\partial^2\mathbb{LL}/\partial\omega^2(\omega=\omega_{IN})} \tag{5.09}$$

The second approach utilises Minimum Variance, unconditionally unbiased, estimates of the residual variances and leads to the recurrence relation of (4.62) in the previous chapter, namely,

$$\omega_{OUT} = \omega_{IN} + \frac{\mathbb{T}_0}{\mathbb{T}_1(\omega_{IN})}.\frac{(\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) - \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0)}{(\mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN}) - \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}))} \tag{5.10}$$

Unfortunately both of the above feedback procedures, (so called because ω_{OUT} is calculated from ω_{IN} and then fed back into the process as ω_{IN} until presumed convergence), have their limitations. However, before examining them in any further detail, it is instructive to review some of the properties of, and related to, their main functional component, $\mathbb{Q}_k$, (of which $\mathbb{T}_k$ is a special case), and, in particular, the ratio $\mathbb{Q}_{k+1}/\mathbb{Q}_k$.

5.211 Properties relating to the functions $\mathbb{Q}_k$ and $\mathbb{Q}_{k+1}/\mathbb{Q}_k$

The function $\mathbb{Q}_k$, or more fully $\mathbb{Q}_k(\omega,\underline{\vartheta}_d,\mathbf{y}_T)$, for k = 0, 1, ... etc. is given by (4.60) of chapter four:

$$\mathbb{Q}_k = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-k}.\mathbf{D}.\mathbf{y}_T \qquad (5.11)$$

where $\mathbf{y}_T$ is the T×1 time series vector consisting of observed values y_T, y_{T-1}, ... , y_1; Ω_ω is the (T-d)×(T-d) matrix, given by equation (4.13) in chapter four as:

$$\Omega_\omega = \omega.\mathbf{I}_{T-d} + \mathbf{D}.\mathbf{D}^T \qquad (5.12)$$

and $\mathbf{D}$ is the (T-d)×T matrix given by equation (3.03) of chapter three as:

$$\mathbf{D} = \begin{bmatrix} 1 , -\vartheta_1, -\vartheta_2 ....., -\vartheta_d, 0 , 0 , 0 , ..., 0 \\ 0 , 1 , -\vartheta_1, -\vartheta_2 ....., -\vartheta_d, 0 , 0 , 0 , .., 0 \\ 0 , 0 , ... \text{etc} \end{bmatrix} \qquad (5.13)$$

PROPERTY 1: The eigenvalues of Ω_ω are given by $\lambda_i = \omega + \xi_i^T.\xi_i$ for i = 1 to T-d with $\xi_i = \mathbf{D}^T.\mathbf{v}_i$, where $\mathbf{v}_i = \mathbf{v}_i(\underline{\vartheta}_d)$ are the eigenvectors and $\Lambda_i = \Lambda_i(\underline{\vartheta}_d) = \xi_i^T.\xi_i$ are the eigenvalues of $\mathbf{D}.\mathbf{D}^T$.

The symmetric invertible matrix $\mathbf{D}.\mathbf{D}^T$ can be written $\mathbf{V}.\mathbf{E}.\mathbf{V}^T$ where $\mathbf{V} = \mathbf{V}(\underline{\vartheta}_d)$ is the matrix whose columns are its eigenvectors, $\mathbf{v}_i$, and $\mathbf{E}$ is the diagonal matrix of its eigenvalues, Λ_i. Therefore $\Lambda_i = \mathbf{v}_i^T.\mathbf{D}.\mathbf{D}^T.\mathbf{v}_i$

$= \xi_i^T.\xi_i$ where $\xi_i = \mathbf{D}^T.\mathbf{v}_i$.

But the matrix Ω_ω can be written $\omega.\mathbf{V}.\mathbf{V}^T + \mathbf{V}.\mathbf{E}.\mathbf{V}^T = \mathbf{V}.(\omega.\mathbf{I}_{T-d} + \mathbf{E}).\mathbf{V}^T$. Therefore $\mathbf{V}$ is also the matrix of the eigenvectors of Ω_ω, and $\omega.\mathbf{I}_{T-d} + \mathbf{E}$ is the diagonal matrix of its eigenvalues, λ_i, i.e. $\lambda_i = \omega + \Lambda_i$.

The important points here are that the eigenvalues are all positive linear functions of ω whose gradients $\partial\lambda_i/\partial\omega$ are all unity, and the eigenvectors, $\mathbf{v}_i = \mathbf{v}_i(\underline{\vartheta}_d)$, are functions of $\underline{\vartheta}_d$ but not of ω.

PROPERTY 2: $\sum_{i=1}^{T-d} \Lambda_i = \sum_{i=1}^{T-d} \xi_i^T.\xi_i = (T-d).\left(1 + \sum_{j=1}^{d} \vartheta_j^2\right)$

Since Λ_i are the eigenvalues of the matrix $\mathbf{D}.\mathbf{D}^T$ defined in property 1, their sum is simply the trace of $\mathbf{D}.\mathbf{D}^T$ i.e. the sum of the elements on its leading diagonal. From inspection of the square (T-d)×(T-d) matrix $\mathbf{D}.\mathbf{D}^T$ every element on its leading diagonal is equal to $1 + \vartheta_1^2 + \vartheta_2^2 + \ldots + \vartheta_d^2$. Hence the result follows.

Note also that from property 1 this gives an equivalent result for the sum of the eigenvalues, λ_i, of Ω_ω, and hence the mean eigenvalue $\bar{\lambda}$,

$$\sum_{i=1}^{T-d} \lambda_i = \sum_{i=1}^{T-d} \left(\omega + \xi_i^T.\xi_i\right) = (T-d).\left(\omega + 1 + \sum_{j=1}^{d} \vartheta_j^2\right)$$

$$\text{i.e. } \bar{\lambda} = \left(\omega + 1 + \sum_{j=1}^{d} \vartheta_j^2\right)$$

PROPERTY 3: $\mathbb{Q}_k = (\xi_1^T.\mathbf{y}_T)^2/\lambda_1^k + (\xi_2^T.\mathbf{y}_T)^2/\lambda_2^k + \ldots. + (\xi_{T-d}^T.\mathbf{y}_T)^2/\lambda_{T-d}^k$ where each vector $\xi_i = \xi_i(\underline{\vartheta}_d) = \mathbf{D}^T.\mathbf{v}_i$, and each $\mathbf{v}_i$ is the eigenvector associated with the eigenvalue Λ_i given by property 1, above.

$$\mathbb{Q}_k = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-k}.\mathbf{D}.\mathbf{y}_T$$

$= \mathbb{TR}[\mathbf{y}_T^T.\mathbf{D}^T.\Omega_\omega^{-k}.\mathbf{D}.\mathbf{y}_T]$ since the trace of a scalar is the scalar itself

$= \mathbb{TR}[\Omega_\omega^{-k}.\mathbf{D}.\mathbf{y}_T.\mathbf{y}_T^T.\mathbf{D}^T]$ using a property of the trace of a matrix

$= \mathbb{TR}[\mathbf{V}.(\omega.\mathbf{I}_{T-d} + \mathbf{E})^{-k}.\mathbf{V}^T.\mathbf{D}.\mathbf{y}_T.\mathbf{y}_T^T.\mathbf{D}^T]$ from property 1

$= \mathbb{TR}[(\omega.\mathbf{I}_{T-d} + \mathbf{E})^{-k}.\mathbf{V}^T.\mathbf{D}.\mathbf{y}_T.\mathbf{y}_T^T.\mathbf{D}^T.\mathbf{V}]$ because it is a trace again

$= \mathbf{v}_1^T.\mathbf{D}.\mathbf{y}_T.\mathbf{y}_T^T.\mathbf{D}^T.\mathbf{v}_1/\lambda_1^k + \ldots + \mathbf{v}_{T-d}^T.\mathbf{D}.\mathbf{y}_T.\mathbf{y}_T^T.\mathbf{D}^T.\mathbf{v}_{T-d}/\lambda_{T-d}^k$

$= (\xi_1^T.\mathbf{y}_T)^2/\lambda_1^k + (\xi_2^T.\mathbf{y}_T)^2/\lambda_2^k + \ldots + (\xi_{T-d}^T.\mathbf{y}_T)^2/\lambda_{T-d}^k$

where $\xi_i = \mathbf{D}^T.\mathbf{v}_i$ and $\lambda_i = \omega + \xi_i^T.\xi_i$ are the diagonal elements of the diagonal matrix $\omega.\mathbf{I}_{T-d} + \mathbf{E}$ from property 1.

Note that when k=0 the above expansion implies that $\mathbb{Q}_0 = \mathbf{y}_T^T.\mathbf{D}^T.\mathbf{D}.\mathbf{y}_T$ i.e. that Ω_ω^0 is defined as $\mathbf{I}_{T-d}$, an identity matrix of dimension T-d.

PROPERTY 4: $\mathbb{Q}_k > 0$, for k = 0, 1, etc.

This follows directly from property 3, since, except for trivial cases, each $\lambda_i > 0$ and each $(\xi_i^T.\mathbf{y}_T)^2 > 0$.

PROPERTY 5: $\mathbb{Q}_k \to 0$, as $\omega \to \infty$.

This follows from properties 1 and 3 since as $\omega \to \infty$, each $\lambda_i \to \infty$ from property 1, and each $(\xi_i^T.\mathbf{y}_T)^2$ in property 3 is not a function of ω.

PROPERTY 6: $\partial\mathbb{Q}_k/\partial\omega = -k.\mathbb{Q}_{k+1} \leq 0$, for k = 1, 2 etc.

$$\partial\mathbb{Q}_k/\partial\omega = \partial/\partial\omega.\left(\sum^i (\xi_i^T.\mathbf{y}_T)^2/(\omega + \xi_i^T.\xi_i)^k\right) \text{ from property 3}$$

$$= -k.\left(\sum^i (\xi_i^T.\mathbf{y}_T)^2/(\omega + \xi_i^T.\xi_i)^{k+1}\right) = -k.\mathbb{Q}_{k+1}$$

which is negative since $\mathbb{Q}_{k+1} > 0$ from property 4.

PROPERTY 7: $\mathbb{Q}_k$ is a strictly decreasing function of ω, for k=1, 2 etc.

This follows from, amongst other things, property 6, since the

gradient of $\mathbb{Q}_k$ is negative.

PROPERTY 8: $\mathbb{T}_k$ is a special case of $\mathbb{Q}_k$ where $\mathbf{y}_T$ is chosen such that $(\xi_i^T.\mathbf{y}_T)^2 = 1$ for i = 1, 2, ... , T-d.

This follows from property 3 and the definition of $T_k = \mathbb{TR}[\Omega_\omega^{-k}] = 1/\lambda_1^k + 1/\lambda_2^k + ... + 1/\lambda_{T-d}^k$.

Note: A value of $\mathbf{y}_T$ which satisfies the above is $\mathbf{y}_T = \mathbf{D}^T.(\mathbf{D}.\mathbf{D}^T)^{-1}.\mathbf{V}.\mathbf{u}$ where $\mathbf{V}$ is the eigenvector matrix defined in property 1 and $\mathbf{u}$ is a (T-d)×1 vector whose elements are all unity.

Also note the implication that $\mathbb{T}_0 = T-d = \mathbb{TR}[\mathbf{I}_{T-d}]$.

PROPERTY 9: $\mathbb{Q}_{k+1}/\mathbb{Q}_k > 0$ for k = 0, 1, etc.

This follows from property 4 since $\mathbb{Q}_{k+1} > 0$ and $\mathbb{Q}_k > 0$.

PROPERTY 10: $\mathbb{Q}_{k+1}/\mathbb{Q}_k \to 0$, as $\omega \to \infty$.

This follows from properties 1 and 3 since for $\omega >> \Lambda_i$, $\lambda_i \to \omega$ for each λ_i and therefore $\mathbb{Q}_k \to \mathbb{Q}_0/\omega^k$. Hence $\mathbb{Q}_{k+1}/\mathbb{Q}_k \to 1/\omega \to 0$ as $\omega \to \infty$.

PROPERTY 11: $\mathbb{Q}_{k+2}/\mathbb{Q}_{k+1} \geq \mathbb{Q}_{k+1}/\mathbb{Q}_k > 0$ for k = 0, 1, etc.

This follows from property 3 since,

$$\mathbb{Q}_{k+2}.\mathbb{Q}_k = \sum^{i}\sum^{j} (\xi_i^T.\mathbf{y}_T)^2/\lambda_i^{k+2}.(\xi_j^T.\mathbf{y}_T)^2/\lambda_j^k$$

$$\text{and } \mathbb{Q}_{k+1}^2 = \sum^{i}\sum^{j} (\xi_i^T.\mathbf{y}_T)^2/\lambda_i^{k+1}.(\xi_j^T.\mathbf{y}_T)^2/\lambda_j^{k+1}$$

$$\text{Hence } \mathbb{Q}_{k+2}.\mathbb{Q}_k - \mathbb{Q}_{k+1}^2 = \sum^{i}\sum^{j} (\xi_i^T.\mathbf{y}_T)^2.(\xi_j^T.\mathbf{y}_T)^2.\lambda_i^{-k}.\lambda_j^{-k}\left(\lambda_i^{-2} - \lambda_i^{-1}.\lambda_j^{-1}\right)$$

$$= \sum^{i<j}\sum^{j} (\xi_i^T.\mathbf{y}_T)^2.(\xi_j^T.\mathbf{y}_T)^2.\lambda_i^{-k}.\lambda_j^{-k}\left(\lambda_i^{-2} + \lambda_j^{-2} - 2.\lambda_i^{-1}.\lambda_j^{-1}\right)$$

$$= \sum^{i<j} \sum^{j} (\xi_i^T.\mathbf{y}_T)^2.(\xi_j^T.\mathbf{y}_T)^2.\lambda_i^{-k}.\lambda_j^{-k}\left(\lambda_i^{-1} - \lambda_j^{-1}\right)^2 \geq 0$$

Note that the equality is only met when all of the eigenvalues are equal which is only true for trivial cases.

$$\text{Therefore } \mathbb{Q}_{k+2}.\mathbb{Q}_k \geq \mathbb{Q}_{k+1}^2$$

Hence $\mathbb{Q}_{k+2}/\mathbb{Q}_{k+1} \geq \mathbb{Q}_{k+1}/\mathbb{Q}_k > 0$ from properties 4 and 9, since $\mathbb{Q}_{k+1} > 0$ and $\mathbb{Q}_k > 0$ and also $\mathbb{Q}_{k+1}/\mathbb{Q}_k > 0$ and again note that this is a strict inequality in all but trivial cases.

Note also that this is one property which is not also necessarily true for the original function $\mathbb{Q}_k$, i.e. $\mathbb{Q}_{k+1} \geq \mathbb{Q}_k$ is <u>not</u> in general true, as the next property shows.

PROPERTY 12: $0 < \mathbb{Q}_{k+p}/\mathbb{Q}_k < 1/\omega^p$.

From properties 2 and 3, we have

$$\omega^p.\mathbb{Q}_{k+p} = \sum^{i} \left((\xi_i^T.\mathbf{y}_T)^2/(\omega + \xi_i^T.\xi_i)^{k+p}\right).\omega^p$$
$$= \sum^{i} \left((\xi_i^T.\mathbf{y}_T)^2/(\omega + \xi_i^T.\xi_i)^{k}\right).\left(\omega/(\omega + \xi_i^T.\xi_i)\right)^p$$

$$< \sum^{i} \left((\xi_i^T.\mathbf{y}_T)^2/(\omega + \xi_i^T.\xi_i)^{k}\right) \text{ since } \omega/(\omega + \xi_i^T.\xi_i) < 1$$

$$\text{Hence } \omega^p.\mathbb{Q}_{k+p} < \mathbb{Q}_k$$

In addition $\mathbb{Q}_k > 0$ from property 4 and the result follows.

PROPERTY 13: $1/\omega > \mathbb{Q}_{k+2}/\mathbb{Q}_{k+1} \geq \mathbb{Q}_{k+1}/\mathbb{Q}_k > 0$ for k = 0, 1, etc.

This follows directly from properties 11 and 12. Note also the special case $(\omega + 1 + \Sigma\, \vartheta_i^2).\mathbb{T}_1 > \mathbb{T}_0$ since $\mathbb{T}_0/\mathbb{T}_1$ is the harmonic mean eigenvalue, which is always less than the arithmetic mean eigenvalue, ($\bar{\lambda} = \omega + 1 + \Sigma\, \vartheta_i^2$, from property 1), for positive eigenvalues. Hence,

$$1/\omega > \mathbb{T}_{k+2}/\mathbb{T}_{k+1} \geq \mathbb{T}_{k+1}/\mathbb{T}_{k} > 1/\bar{\lambda} \text{ for } k = 0,\ 1,\ \text{etc.}$$

PROPERTY 14: $\partial(\mathbb{Q}_{k+1}/\mathbb{Q}_{k})/\partial\omega = -\mathbb{Q}_{k+1}/\mathbb{Q}_{k}(\mathbb{Q}_{k+2}/\mathbb{Q}_{k+1} + k(\mathbb{Q}_{k+2}/\mathbb{Q}_{k+1} - \mathbb{Q}_{k+1}/\mathbb{Q}_{k})) \leq 0$ for k = 0, 1, 2 etc.

This follows from straight differentiation using property 6 and properties 9 and 11 which ensure that the differential is negative, and only zero when $\mathbb{Q}_{k+1}$ and $\mathbb{Q}_{k+2}$ are zero.

PROPERTY 15: $\mathbb{Q}_{k+1}/\mathbb{Q}_{k}$ is a decreasing function of ω, for k = 0, 1, 2 etc.

This follows from property 14 since its gradient is negative for all non-trivial cases.

PROPERTY 16: $\mathbb{Q}_{k+1}/\mathbb{Q}_{k}$ is finite for all values of ω.

Because of property 4, $\mathbb{Q}_{k+1}/\mathbb{Q}_{k}$ is the ratio of two positive values which must also be positive and finite. Note also that because of property 15 its maximum is at $\omega = 0$.

5.212 The Constant and Linear Models

In order to clarify the procedures in the remainder of this chapter, I have chosen two simple models as examples. Both models utilise the same time series, $\mathbf{y}_T$, namely that of figure 1.1 in chapter one.

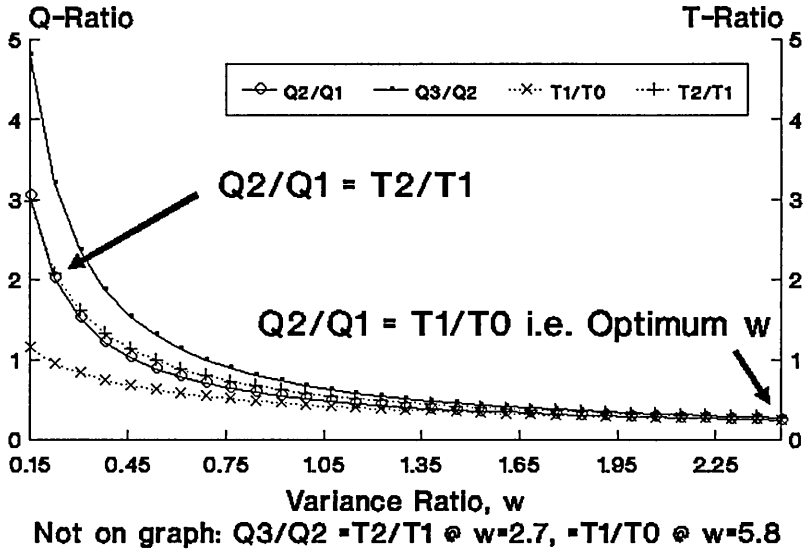
The "constant" model, (also referred to as the basic model of section **3.3**, equation 3.28), employs the structural, or trend, equation, $x_t = x_{t-1} + a_t$ i.e. it has lag d=1 and autoregressive parameter $\vartheta_1 = +1$ and is so-called because of its invariance to trends with a constant mean, (see sections **1.3** and **3.5**).

The "linear" model employs the trend equation $x_t = 2.x_{t-1} - x_{t-2} + a_t$ i.e. it has lag d=2 and autoregressive parameters $\vartheta_1 = +2$, $\vartheta_2 = -1$ and is

invariant to trends which are linear over time.

Figure 5.1 shows examples of the $\mathbb{Q}$-ratio, $(\mathbb{Q}_{k+1}/\mathbb{Q}_k)$, family of curves for k = 1 and 2 for the constant model. Note that, as previously demonstrated, they are positive, finite, non-intersecting curves such that $\mathbb{Q}_{k+2}/\mathbb{Q}_{k+1} > \mathbb{Q}_{k+1}/\mathbb{Q}_k$, which tend to zero as ω tends to infinity.

Q-Ratio Curves (Constant Model)
Figure 5.1



Not on graph: Q3/Q2 =T2/T1 @ w=2.7, =T1/T0 @ w=5.8

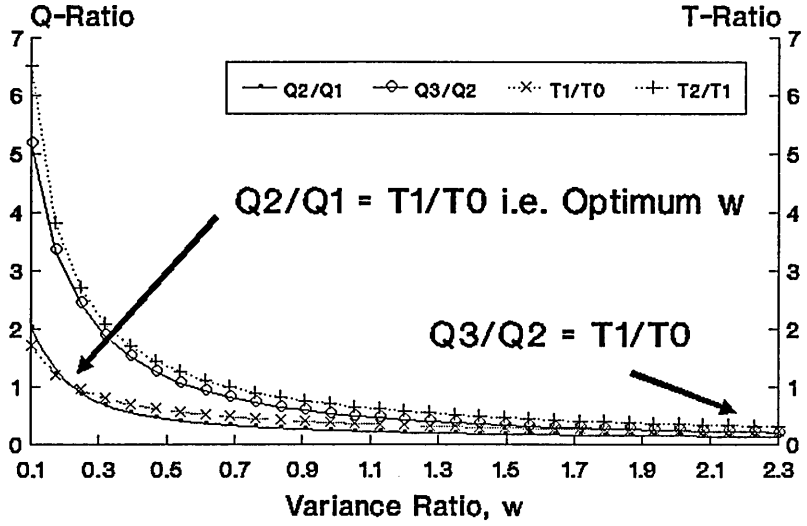
Figure 5.1 also shows $\mathbb{T}$-ratios, (see property 8 above), for k=0 and 1, for the same model which also form another family with the same characteristics as the $\mathbb{Q}$-ratios.

Note that both $\mathbb{Q}$-ratios intersect both $\mathbb{T}$-ratios once only, (excluding the case $\omega=\infty$ of course where all ratios eventually meet), the most important intersection being when $\mathbb{Q}_2/\mathbb{Q}_1 = \mathbb{T}_1/\mathbb{T}_0$ which solves equation 5.05 and hence gives an optimum value of ω.

However it is seen that this pattern of intersections breaks down when we employ a linear model as in figure 5.2. For this model we see that the $\mathbb{T}$-ratio $\mathbb{T}_2/\mathbb{T}_1$ does not intersect either of the $\mathbb{Q}$-ratios and remains at a higher level than both for all values of ω.

Hence, even when using the same time series data it does not seem possible to be able to predict how the $\mathbb{Q}$ and $\mathbb{T}$ ratios will relate to each other.

Q-Ratio Curves (Linear Model)
Figure 5.2



Experience of investigating different data sets and different models have resulted in most conjectures about the curves being rejected. The only conjecture which does seem to hold for every example I have looked at so far is that "there can only be, at most, one intersection between any particular $\mathbb{Q}$-ratio and any particular $\mathbb{T}$-ratio", although this, as yet, remains unproven.

However, even though our knowledge of how these curves behave is far from complete, we can, nevertheless, use the information we have obtained so far to some extent in solving equation 5.05 using the two approaches of equations 5.09 and 5.10.

5.213 The Likelihood Approach

Applying property 13 to equation (5.08), we have,

$$\partial^2 LL/\partial\omega^2 = -T_0/2.(2.Q_3(\omega)/Q_1(\omega) - Q_2^2(\omega)/Q_1^2(\omega) - T_2(\omega)/T_0) \qquad (5.14)$$

and hence using equations (5.08) and (5.14) equation (5.09) becomes:

$$\omega_{OUT} = \omega_{IN} + \frac{(Q_2(\omega_{IN})/Q_1(\omega_{IN}) - T_1(\omega_{IN})/T_0)}{(2.Q_3(\omega_{IN})/Q_1(\omega_{IN}) - Q_2^2(\omega_{IN})/Q_1^2(\omega_{IN}) - T_2(\omega_{IN})/T_0)} \qquad (5.15)$$

Maximum Likelihood Estimation of w
Figure 5.3

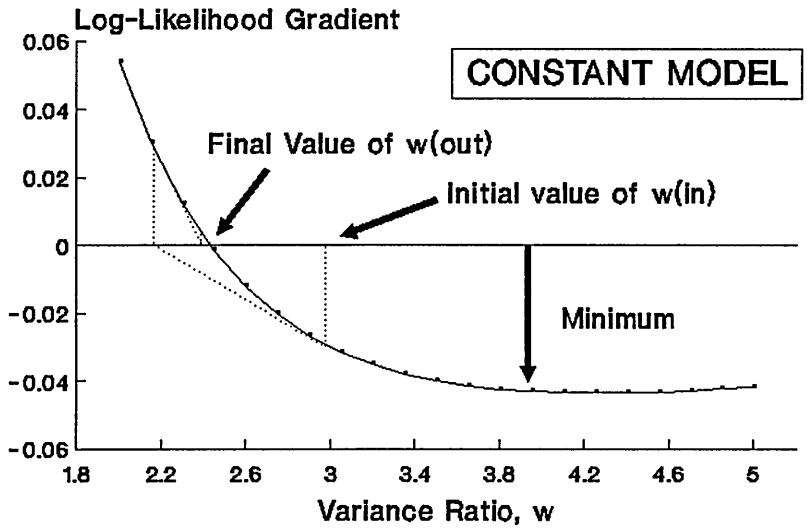


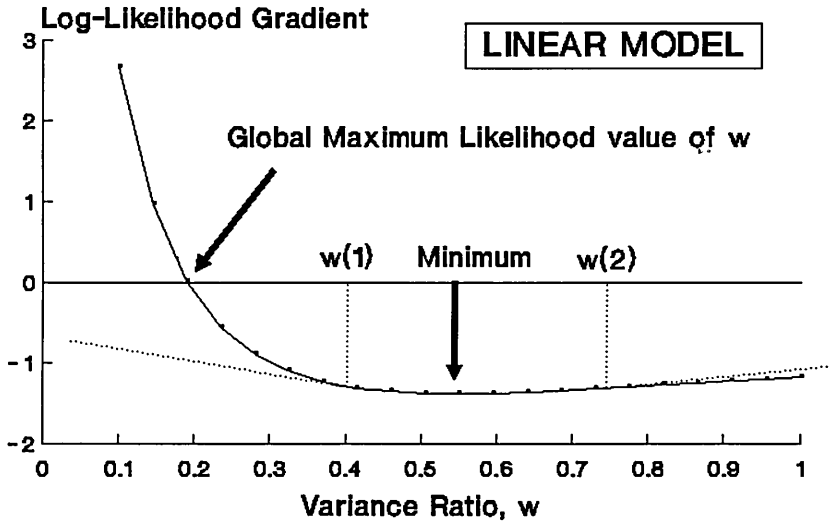
Figure 5.3 shows a plot of the gradient of the log-likelihood equation, $\partial LL/\partial\omega$, for the constant model to demonstrate how equation (5.15) operates. From an initial ω value, ω_{IN}, the value of ω_{OUT} is found as the point at which the tangent to $\partial LL/\partial\omega(\omega=\omega_{IN})$ meets the ω axis. This is then set as the new ω_{IN} value and the process repeated until convergence at the final ω_{OUT} value, which occurs where the log-likelihood gradient crosses the ω axis.

The log-likelihood gradient in figure 5.3 is well-behaved. From an initial positive value it crosses the ω axis, reaches a minimum and then, (although not shown), tends back to zero as ω approaches infinity. Since the curve crosses the ω axis from a positive to a

negative value, this intersection is at a maximum of the log-likelihood, (and hence likelihood), function and the value of ω at the intersection is its local maximum likelihood value. Also since this is the curve's only intersection with the ω axis it must also be its global maximum.

Figure 5.4 shows a similar situation for the linear model. Again the log-likelihood gradient is well-behaved and from a suitably chosen ω_{IN} will converge to a global maximum by successive applications of equation 5.15.

Maximum Likelihood Estimation of w
Figure 5.4

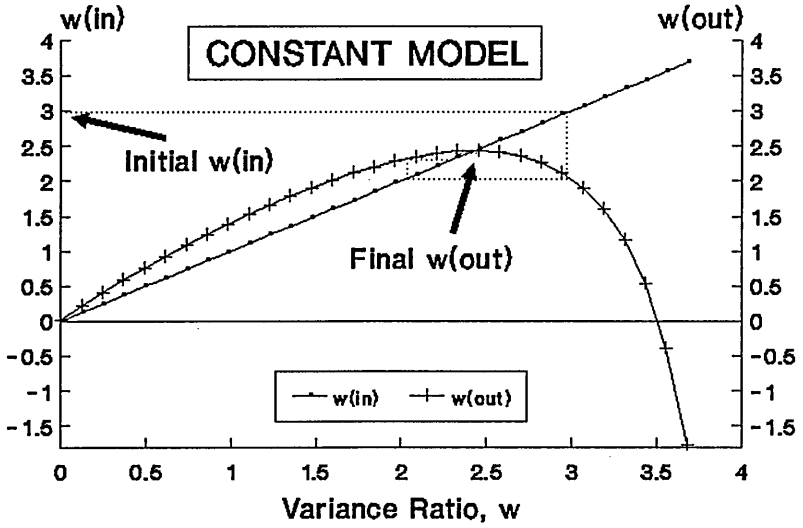


Unfortunately, there is nothing to say that the log-likelihood gradient will always be so well-behaved. For instance, there is no proof that there will always be a local maximum, let alone a global one. Moreover even if the log-likelihood gradient is well-behaved there are still problems with applying equation 5.15.

Firstly, if ω_{IN} is chosen too close to the log-likelihood gradient's minimum, such as ω_1 in figure 5.4, the application of equation 5.15 may result in a negative value of ω_{OUT} and the recursion will break down.

Secondly, if ω_{IN} is chosen to be a value which is higher than the value of ω at which the minimum of the log-likelihood gradient occurs, such as ω_2 in figure 5.4, it should be apparent that the application of equation 5.15 will result in higher and higher values of ω which will direct the process to infinity.

Maximum Likelihood Estimation of w
Figure 5.5



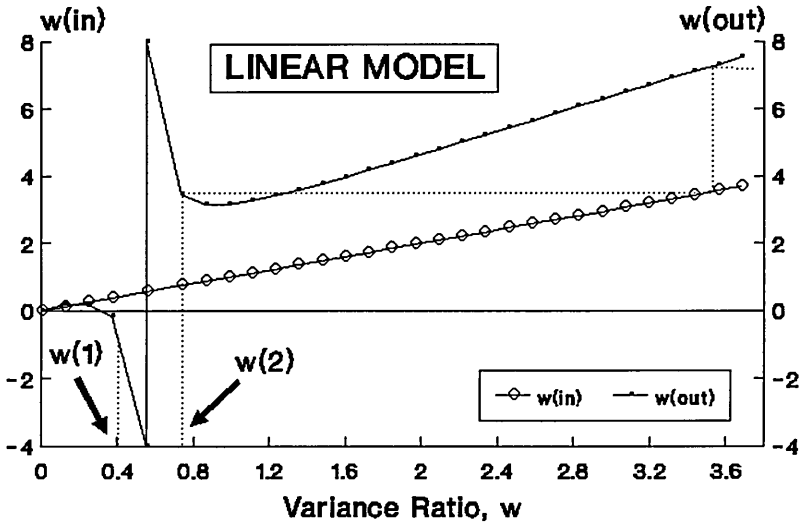
An alternative way of looking at things is given by plotting an ω_{IN}/ω_{OUT} graph as in figures 5.5 and 5.6. Figure 5.5 shows the ω_{IN}/ω_{OUT} equivalent graph for the constant model of figure 5.3, albeit only for values of ω up to about 3.8, which is the minimum point of the log-likelihood gradient. The graph recreates the $\omega_{IN} \rightarrow \omega_{OUT}$ stage by moving vertically from the ω_{IN} line to the ω_{OUT} curve, and sets $\omega_{OUT}=\omega_{IN}$ by moving horizontally from the ω_{OUT} curve to the ω_{IN} line. (Note the ω_{IN} line is just the straight line $\omega_{IN}=\omega_{OUT}$). From the graph we can immediately see that for values of ω up to about 3.5, i.e. the point where the ω_{OUT} curve crosses the ω axis, this procedure will always result in convergence.

Figure 5.6 gives a wider picture of the ω_{IN}/ω_{OUT} equivalent graph for the linear model of figure 5.4, which now includes the the minimum point of the log-likelihood gradient at about ω=0.5, which manifests

itself as a singularity.

Again it is easy to see from the figure that initial values for ω_{IN} such as ω_1 will always produce negative values for ω_{OUT}, and values such as ω_2 will always result in a process which iterates to infinity.

Maximum Likelihood Estimation of w
Figure 5.6



What does all this tell us about using equation (5.15) to find the global optimum value of ω. In general very little, since there are so many if's and but's regarding which form the log-likelihood gradient function may take that even to be certain of obtaining a local optimum is not definite and may not even exist.

However, as long as the log-likelihood gradient is "well-behaved", it does suggest choosing a small value for the initial value of ω_{IN} to locate at least one of the possible optima.

Also it suggests that we can expect to get negative values of ω_{OUT} if ω_{IN} is large, and to beware of infinite recursions when ω_{IN} is very large. Finally, inspection of figures 5.1 and 5.2, suggest choosing an initial values of ω_{IN} such that $\mathbb{Q}_2/\mathbb{Q}_1$ is greater than $\mathbb{T}_1/\mathbb{T}_0$.

Suppose we now rearrange equation (5.15), (dropping the parameter ω_{IN} which is understood), we get:

$$w_{OUT} = w_{IN} + \frac{\mathbb{Q}_2/\mathbb{Q}_1 - \mathbb{T}_1/\mathbb{T}_0}{(\mathbb{Q}_2/\mathbb{Q}_1).(\mathbb{Q}_3/\mathbb{Q}_2 - \mathbb{Q}_2/\mathbb{Q}_1) + (\mathbb{Q}_2/\mathbb{Q}_1).(\mathbb{Q}_3/\mathbb{Q}_2) - (\mathbb{T}_2/\mathbb{T}_1).(\mathbb{T}_1/\mathbb{T}_0)}$$

....(5.16)

If ω_{IN} is chosen such that $\mathbb{Q}_2/\mathbb{Q}_1 > \mathbb{T}_2/\mathbb{T}_1$, then, since $\mathbb{Q}_3/\mathbb{Q}_2 > \mathbb{Q}_2/\mathbb{Q}_1$ and $\mathbb{T}_2/\mathbb{T}_1 > \mathbb{T}_1/\mathbb{T}_0$ from Q-ratio properties 8 and 11, this means that $\mathbb{Q}_3/\mathbb{Q}_2 > \mathbb{Q}_2/\mathbb{Q}_1 > \mathbb{T}_2/\mathbb{T}_1 > \mathbb{T}_1/\mathbb{T}_0$ and hence, because of properties 8 and 9, that $(\mathbb{Q}_2/\mathbb{Q}_1).(\mathbb{Q}_3/\mathbb{Q}_2) > (\mathbb{T}_2/\mathbb{T}_1).(\mathbb{T}_1/\mathbb{T}_0)$.

From this we can conclude that the fractional addition to ω_{IN} in equation (5.16) will always be positive, and hence ω_{OUT} will always be positive if ω_{IN} is such that $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) > \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$.

Note that this does not, in itself, show that successive recursions of equation (5.15) will also be valid since the condition $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) > \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$ does not necessarily imply that $\mathbb{Q}_2(\omega_{OUT})/\mathbb{Q}_1(\omega_{OUT}) > \mathbb{T}_2(\omega_{OUT})/\mathbb{T}_1(\omega_{OUT})$ for $\omega_{OUT} > \omega_{IN}$ and, even if it did it, does not exclude the possibility of an infinite recursion.

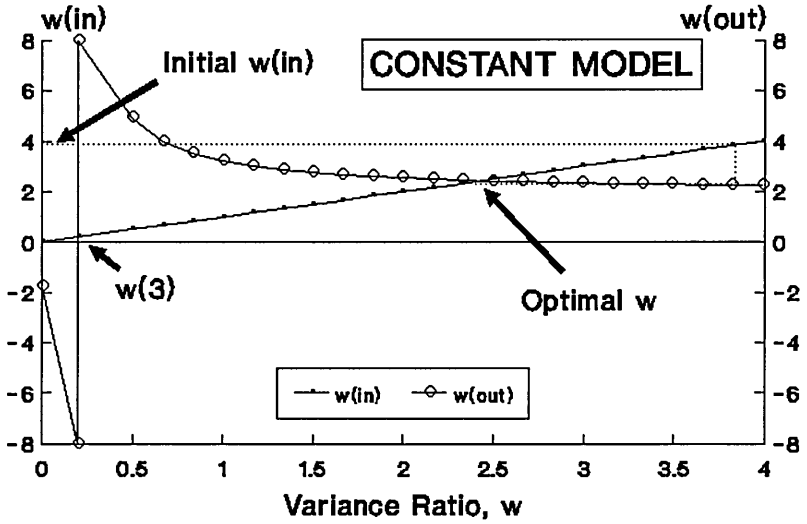
However it does mean that at least <u>one</u> step of equation (5.15) can be validly performed, which we will find is something which can be utilised when taken in conjunction with the second approach to finding the optimal value of ω using equation (5.10).

5.214 The Minimum Variance Approach

This approach utilises the recurrence relation of equation (5.10), the ω_{IN}/ω_{OUT} graphs for which are shown in figures 5.7 and 5.8 for the constant and linear models respectfully. It is apparent that the characteristics of these graphs are quite different from those of figures 5.5 and 5.6 which utilise the maximum likelihood method.

Figure 5.7 shows that no values of ω_{IN} below the singularity at ω_3, i.e. at about ω=0.2 will result in the convergence of equation (5.10) as they all lead instantly to a negative value for ω_{OUT}. (Note that the value $\omega=\omega_3$ is the point at which $\mathbb{Q}_2(\omega_3)/\mathbb{Q}_1(\omega_3)=\mathbb{T}_2(\omega_3)/\mathbb{T}_1(\omega_3)$ in both equation (5.10) and figure 5.1).

Minimum Variance Estimation of w
Figure 5.7



Alternatively, for any value of ω_{IN} greater than ω_3, equation (5.10) <u>always</u> leads to convergence at the optimal ω, since the w_{OUT} flattens out at a level above ω_3 at large values of ω.

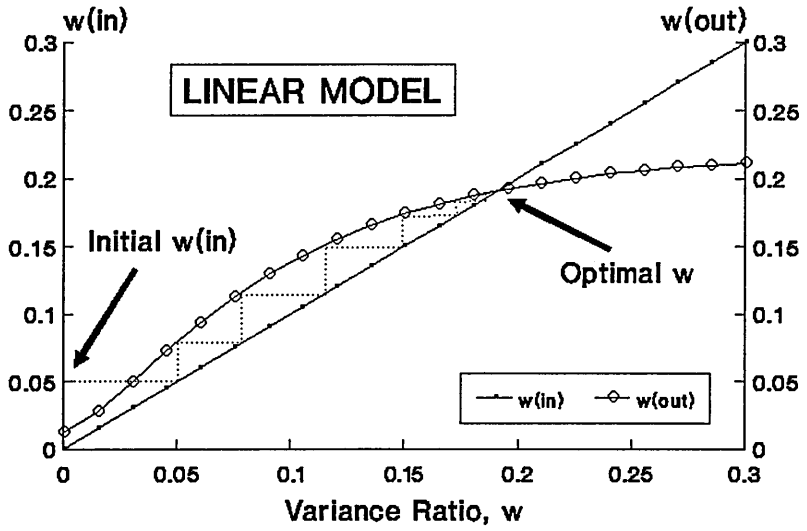
Unlike the likelihood example of figure 5.5, there are no values of ω_{IN} which result in an infinite recursion.

The contrast between the two approaches is even more marked when it comes to the linear model shown in figure 5.8. For this model <u>every</u> value of ω_{IN} leads to the optimum ω, since there are no positive values of ω which will satisfy the singularity condition, $\mathbb{Q}_2(\omega)/\mathbb{Q}_1(\omega)=\mathbb{T}_2(\omega)/\mathbb{T}_1(\omega)$.

In fact for larger values of ω than those shown on the graph, the ω_{OUT} curve takes on a maximum, followed by a minimum, followed by a trend

to infinity, which although interesting is completely irrelevant as regards the convergence properties of equation (5.10) since it never again either crosses the ω_{IN} line or becomes negative which are the crucial determining factors.

Minimum Variance Estimation of w
Figure 5.8



We can obtain a little more insight into what is happening by reflecting again on how equation (5.10) was derived.

It was produced from equations (4.62) and (4.65) of chapter four, where ω_{OUT} is defined as the ratio of the minimum variance, unconditionally unbiased, residual variance estimates $\hat{\sigma}^2_a(\omega_{IN})$ and $\hat{\sigma}^2_e(\omega_{IN})$ i.e.

$$\omega_{OUT} = \hat{\sigma}^2_a(\omega_{IN})/\hat{\sigma}^2_e(\omega_{IN}) \qquad (5.17)$$

where the measurement variance estimate, $\hat{\sigma}^2_e(\omega_{IN})$, was given by equation (4.61) of chapter four, i.e.

$$\hat{\sigma}^2_e(\omega_{IN}) = \frac{\mathbb{T}_2.\mathbb{Q}_1 - \mathbb{T}_1.\mathbb{Q}_2}{\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1.\mathbb{T}_1} = (\mathbb{Q}_1/\mathbb{T}_0).\frac{(\mathbb{T}_2/\mathbb{T}_1 - \mathbb{Q}_2/\mathbb{Q}_1)}{(\mathbb{T}_2/\mathbb{T}_1 - \mathbb{T}_1/\mathbb{T}_0)} \qquad (5.18)$$

and the structural variance estimate, $\hat{\sigma}^2_a(\omega_{IN})$, was given by equation (4.59) of chapter four, i.e.

$$\hat{\sigma}^2_a(\omega_{IN}) = \omega_{IN}.\hat{\sigma}^2_e(\omega_{IN}) + \frac{\mathbb{T}_0.\mathbb{Q}_2 - \mathbb{T}_1.\mathbb{Q}_1}{\mathbb{T}_0.\mathbb{T}_2 - \mathbb{T}_1.\mathbb{T}_1}$$

$$= \omega_{IN}.\hat{\sigma}^2_e(\omega_{IN}) + (\mathbb{Q}_1/\mathbb{T}_1).\frac{(\mathbb{Q}_2/\mathbb{Q}_1 - \mathbb{T}_1/\mathbb{T}_0)}{(\mathbb{T}_2/\mathbb{T}_1 - \mathbb{T}_1/\mathbb{T}_0)} \qquad (5.19)$$

Immediately we see that if equation (5.17) is to be a valid recursion the two residual variance estimates should both be positive.

From equation (5.18), we see that for $\hat{\sigma}^2_e(\omega_{IN}) > 0$, we require $\mathbb{Q}_2/\mathbb{Q}_1 < \mathbb{T}_2/\mathbb{T}_1$, since from $\mathbb{Q}$-ratio properties 4 and 11, $\mathbb{T}_1/\mathbb{T}_0 < \mathbb{T}_2/\mathbb{T}_1$ and $\mathbb{Q}_1/\mathbb{T}_0 > 0$.

Also, from equation (5.19), using the same reasoning, we see that if $\hat{\sigma}^2_e(\omega_{IN}) > 0$, then $\hat{\sigma}^2_a(\omega_{IN})$ will also be positive if $\mathbb{Q}_2/\mathbb{Q}_1 > \mathbb{T}_1/\mathbb{T}_0$.

Hence a necessary condition for a valid step in the recursion of equation (5.17) is that $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$, and a sufficient condition is that $\mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN}) < \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$.

5.215 A Joint Approach

We now begin to see how this approach might be used in conjunction with the likelihood approach where we found that a sufficient condition for a valid step in its recursion using equation (5.15) was that $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) > \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$. Hence as long as $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) > \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$ we can use one or the other of equations (5.15) and (5.17) as the next step in the recursion and be sure it will be valid.

This leaves us with case when $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$. Even

though recursions (5.15) and (5.17) do not guarantee a valid step in this situation, we can still apply either of them as long as they are seen to produce a value for ω_{OUT} which is not negative. If we are given a choice, the use of (5.17) is much more desirable than (5.15) for two reasons.

Firstly, as we have seen (5.15) tends to produce an infinite recursion in this situation, whereas this is not only unlikely, but is actually impossible if (5.17) is used.

If we examine (5.17) or in its full form (5.10), we see that it can be written in the form,

$$\omega_{OUT} = \omega_{IN} + \delta(\omega_{IN}) \tag{5.20}$$

$$\text{where} \quad \delta(\omega_{IN}) = \frac{\mathbb{T}_0}{\mathbb{T}_1(\omega_{IN})} \cdot \frac{(\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) - \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0)}{(\mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN}) - \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}))} \tag{5.21}$$

As we have seen, if $\mathbb{Q}_2/\mathbb{Q}_1 < \mathbb{T}_1/\mathbb{T}_0$, then $\mathbb{Q}_2/\mathbb{Q}_1 < \mathbb{T}_2/\mathbb{T}_1$, since $\mathbb{T}_2/\mathbb{T}_1 > \mathbb{T}_1/\mathbb{T}_0$, and so $\delta(\omega_{IN}) < 0$ and hence ω_{OUT} must always be less than ω_{IN} and therefore cannot recurse to infinity, (although it can recurse to a non- positive value).

Secondly, any optimal value of ω found using a valid recursive step of (5.17) must maximise the likelihood function.

Examination of equation (5.21) shows that,

(i) If $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$

then $\delta(\omega_{IN}) < 0$

since $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN}) < \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$

$$\text{(ii) If } \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN}) < \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN})$$

$$\text{then } \delta(\omega_{IN}) > 0$$

$$\text{(iii) If } \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN}) < \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN})$$

$$\text{then } \delta(\omega_{IN}) < 0$$

$$\text{since } \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN}) < \mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN}) < \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN})$$

Of these three possible cases only the first two would result from the application of equation (5.17) since if (iii) were true we have already seen that we would apply the likelihood recursion of (5.15).

Hence if $\omega_{OPT} = \omega_{IN} + \delta(\omega_{IN})$ is the final step in application of the recursion of (5.20), and if $0 < \omega_{OPT} < \infty$, then,

if $\omega_{IN} > \omega_{OPT}$, $\delta(\omega_{IN}) < 0$ and $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$, and hence, from equation (5.08), $\partial\mathbb{LL}/\partial\omega(\omega_{IN}) < 0$.

Alternatively if $\omega_{IN} < \omega_{OPT}$, $\delta(\omega_{IN}) > 0$ and $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) > \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$, and, from equation (5.08), $\partial\mathbb{LL}/\partial\omega(\omega_{IN}) > 0$.

Since the size of $\delta(\omega_{IN})$ can be made as small as we choose, then within this tolerance the conditions $\partial\mathbb{LL}/\partial\omega(\omega_{IN}) < 0$ for $\omega_{IN} > \omega_{OPT}$ and $\partial\mathbb{LL}/\partial\omega(\omega_{IN}) > 0$ for $\omega_{IN} < \omega_{OPT}$ are exactly those needed for ω_{OPT} to maximise the log-likelihood, ($\mathbb{LL}$), and hence the likelihood.

Finally we note that by choosing $\delta(\omega_{IN})$ to be sufficiently small, we can always ensure that the final step in the recursion uses equation (5.20) rather than (5.17), since if ω_{IN} and ω_{OPT} are such that $\mathbb{T}_2(\omega_{IN})/\mathbb{T}_1(\omega_{IN}) < \mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN})$ and $\mathbb{T}_1(\omega_{OPT})/\mathbb{T}_0(\omega_{OPT}) = \mathbb{Q}_2(\omega_{OPT})/\mathbb{Q}_1(\omega_{OPT})$, then there will always be a value ω_{OUT} between ω_{IN} and ω_{OPT} such that $\mathbb{T}_1(\omega_{OUT})/\mathbb{T}_0(\omega_{OUT}) < \mathbb{Q}_2(\omega_{OUT})/\mathbb{Q}_1(\omega_{OUT}) < \mathbb{T}_2(\omega_{OUT})/\mathbb{T}_1(\omega_{OUT})$.

5.216 The Slow but Stable Approach

The "joint approach" just described in section **5.215** still leaves us with us with one eventually which we have not made provision for; namely the situation when $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$ and the application of (5.20) would result in a value for ω_{OUT} which was negative.

In this event we offer the following fall-back mechanism which might be termed the "Slow but Stable" approach. We suggest the following recursive equation,

$$\omega_{OUT} = \omega_{IN} + \omega_{IN}^2.(\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) - \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})) \qquad (5.22)$$

for the following reasons.

We can immediately see from the simple, non-fractional form of (5.22) that there will be no problems with regard to singularities.

Also, should (5.22) converge it must converge to value ω_{OPT} such that $\mathbb{Q}_2(\omega_{OPT})/\mathbb{Q}_1(\omega_{OPT}) = \mathbb{T}_1(\omega_{OPT})/\mathbb{T}_0(\omega_{OPT})$ which from (5.08) satisfies the condition $\partial\mathbb{LL}/\partial\omega(\omega_{OPT}) = 0$ and so optimises the likelihood function.

In addition if (5.22) converges to a final step of ω_{OPT} from ω_{IN}, then for $\omega_{IN} < \omega_{OPT}$, $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) > \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$ and hence from (5.08), $\partial\mathbb{LL}/\partial\omega(\omega_{IN}) > 0$. Similarly if $\omega_{IN} > \omega_{OPT}$, $\partial\mathbb{LL}/\partial\omega(\omega_{IN}) < 0$ and hence if the recursion does converge to a value of ω_{OPT} such that $0 < \omega_{OPT} < \infty$, it can only be to a value which maximises the likelihood function.

Figure 5.9 shows the ω_{IN}/ω_{OUT} graph for the "Slow but Steady" approach and the constant model. Two features are immediately apparent.

Firstly, because the ω_{IN} and ω_{OUT} plots are very close together, recursions tend to be very slow, which is the reason why it is not offered on its own as a general recursive algorithm for choosing an optimal ω.

Secondly, the function for ω_{OUT} is never negative. This is no coincidence because,

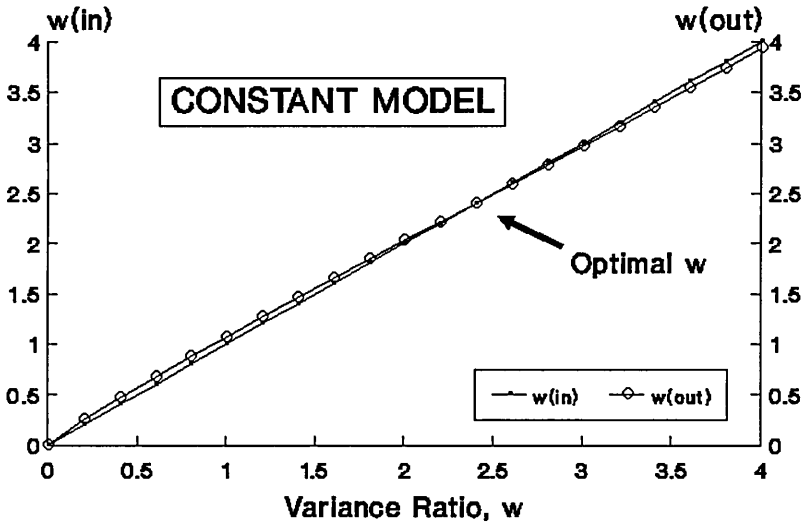
$$-1/\omega < -\mathbb{T}_1/\mathbb{T}_0 < \mathbb{Q}_2/\mathbb{Q}_1 - \mathbb{T}_1/\mathbb{T}_0 < \mathbb{Q}_2/\mathbb{Q}_1 < +1/\omega \qquad (5.23)$$

since both $\mathbb{Q}_2/\mathbb{Q}_1$ and $\mathbb{T}_1/\mathbb{T}_0$ are positive from $\mathbb{Q}$-ratio properties 8 and 9 and the outer limits are given by property 12. Hence,

$$-\omega < \omega^2.(\mathbb{Q}_2/\mathbb{Q}_1 - \mathbb{T}_1/\mathbb{T}_0) < +\omega \qquad (5.24)$$

which ensures that $\omega_{OUT} > \omega_{IN}$ in equation (5.22).

Slow but Stable Estimation of w
Figure 5.9



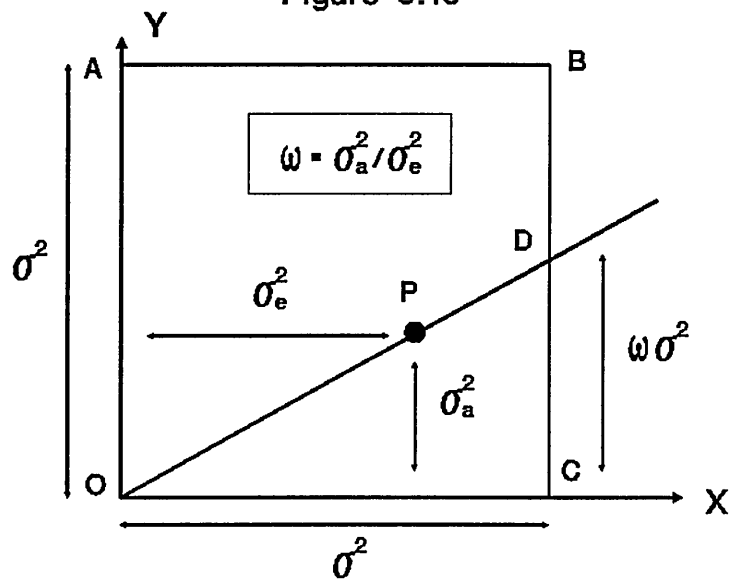
Therefore for situations where $\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) < \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN})$ and the application of (5.20) results in a value for ω_{OUT} which is negative, we can apply equation (5.22) and be certain that it will always result in a positive value for ω_{OUT} and hence not result in the overall recursion breaking down.

5.217 The Choice of a Random Ratio

We are now almost at the stage where we can combine all the features which we have discussed so far into a general algorithm for the estimation of an optimal value for ω. However, because the algorithmn can only find, at most, only one local optimum value of ω for any particular starting value, it will be necessary to repeat it for several different starting values to be reasonably sure that the model's global optimum, if it exists, has been found. We therefore require a mechanism for selecting such a set of starting values.

Because ω is the ratio of two variances, (i.e. $\omega = \sigma_a^2/\sigma_e^2$), either of which can take any value in the range zero to infinity, it would be useful to know the what distribution of ω is when the variances are chosen randomly. Since we have no prior knowledge of what variance values are most likely, this would suggest every variance value between zero and infinity should be treated as equiprobable. Whilst this makes no practical sense for the individual variance distributions themselves, it does lead to an easily interpreted distribution for their ratio.

Random Choice of the Variance Ratio
Figure 5.10



Suppose that σ_a^2 and σ_e^2 are each independently chosen in the range 0 to σ^2. As shown in figure 5.10, this choice can be represented by a point P, which has been randomly chosen within a square OABC of side σ^2 in the X-Y plane.

Also using figure 5.10 it is seen that the probability that any other randomly chosen point, (X,Y), within the square, has a ratio, Y/X, less than or equal to $\omega = \sigma_a^2/\sigma_e^2$ is, (for $0 \leq \omega \leq 1$), equal to the ratio of the area of triangle ODC, (i.e. $\omega.\sigma^4/2$), to that of the square OABC, (i.e. σ^4). Hence $P(Y/X \leq \omega) = \omega/2$, which is not dependent on σ^2.

Similarly, for $1 \leq \omega \leq \infty$, $P(Y/X \leq \omega) = 1 - 1/(2\omega)$, which is also independent of σ^2, and is hence true for all non-negative values of σ^2, however large these might be.

Hence an algorithm for choosing a starting value for ω, i.e. ω_0, is as follows.

1. Choose a value of P in the range $0 \leq P < 1$.

2. If $P < 0.5$ then $\omega_0 = 2P$. If $P > 0.5$ then $\omega_0 = 1/(2-2P)$.

Note: Whether the n sample values of P, i.e P_i for i=1 to n, should be chosen randomly or in some form of even spread over their range, such as $P_i = i/(n+1)$ or $(i-0.5)/n$, is left as an option.

5.22 A GENERAL ALGORITHM

The following steps give a general algorithm for calculating optimum, maximum likelihood values for the variance ratio, ω_{OPT}, and the two residual variances $\sigma_e^2(OPT)$ and $\sigma_a^2(OPT)$.

In preparation we need to choose a sample size, N, starting values $\omega_0(n)$ for n = 1 to N (from section **5.217**, a lower bound, (ω_{LOWER}), upper bound, (ω_{UPPER}) and tolerance, (δ) for ω_{OPT} and finally set n=0 and the log-likelihood value $LL(\omega_{OPT})$ to a large negative lower bound.

1. Set n = n + 1, and ω_{OUT} = starting value, $\omega_0(n)$

2. Set $\omega_{IN} = \omega_{OUT}$

3. Calculate $\mathbb{Q}_2/\mathbb{Q}_1$ and $\mathbb{T}_2/\mathbb{T}_1$ for $\omega=\omega_{IN}$ and hence apply either 4(a) or 4(b)

4(a) If $\mathbb{T}_2/\mathbb{T}_1 < \mathbb{Q}_2/\mathbb{Q}_1$ calculate ω_{OUT} using equation (5.15), i.e.

$$w_{OUT} = w_{IN} + \frac{(\mathbb{Q}_2(w_{IN})/\mathbb{Q}_1(w_{IN}) - \mathbb{T}_1(w_{IN})/\mathbb{T}_0)}{(2.\mathbb{Q}_3(w_{IN})/\mathbb{Q}_1(w_{IN}) - \mathbb{Q}_2^2(w_{IN})/\mathbb{Q}_1^2(w_{IN}) - \mathbb{T}_2(w_{IN})/\mathbb{T}_0)}$$

4(b) If $\mathbb{Q}_2/\mathbb{Q}_1 < \mathbb{T}_2/\mathbb{T}_1$ calculate ω_{OUT} using equations (5.20) and (5.21), i.e.

$$w_{OUT} = w_{IN} + \frac{\mathbb{T}_0}{\mathbb{T}_1(w_{IN})}.\frac{(\mathbb{Q}_2(w_{IN})/\mathbb{Q}_1(w_{IN}) - \mathbb{T}_1(w_{IN})/\mathbb{T}_0)}{(\mathbb{T}_2(w_{IN})/\mathbb{T}_1(w_{IN}) - \mathbb{Q}_2(w_{IN})/\mathbb{Q}_1(w_{IN}))}$$

If $\omega_{OUT} < \omega_{LOWER}$, recalculate ω_{OUT} using equation (5.22), i.e.

$$\omega_{OUT} = \omega_{IN} + \omega_{IN}^2.(\mathbb{Q}_2(\omega_{IN})/\mathbb{Q}_1(\omega_{IN}) - \mathbb{T}_1(\omega_{IN})/\mathbb{T}_0(\omega_{IN}))$$

5. If $|\omega_{OUT} - \omega_{IN}|$ < tolerance, δ, GOTO step 8

6. If $\omega_{OUT} > \omega_{UPPER}$ OR $\omega_{OUT} < \omega_{LOWER}$, GOTO step 9

7. GOTO step 2

8. Calculate $\mathbb{LL}(\omega_{OUT})$ using equations (4.14) and (4.15) of chapter four, i.e.

$$\mathbb{LL}(\omega_{OUT}) = -1/2.\left(\mathbb{T}_0.(1 + \ln(2\pi) - \ln(\mathbb{T}_0)) + \mathbb{T}_0.\ln(\mathbb{Q}_1) + \ln(|\Omega_\omega|\right)$$

and then if $\mathbb{LL}(\omega_{OUT}) > \mathbb{LL}(\omega_{OPT})$ set

$$\omega_{OPT} = \omega_{OUT},\ \sigma_e^2(OPT) = \mathbb{Q}_1/\mathbb{T}_0 \text{ and } \sigma_a^2(OPT) = \omega_{OPT}.\sigma_e^2(OPT)$$

9. If n=N then terminate algorithm, otherwise GOTO step 1

SEASONALITY

In this chapter we extend the estimating procedures of the previous chapters to deal with the case of seasonal time series. This extension is a natural generalisation of the non-seasonal model and easily accommodated within its framework. As such it also serves as a useful summary of the main results covered so far.

6.1 WHITTAKER'S FORMULATION

6.11 FIDELITY

The time series values, y_t, are now decomposed to include a seasonal component, s_t, as well as the previous trend, x_t, and residual fidelity, e_t, components, i.e.

$$y_t = x_t + s_t + e_t \qquad (6.01)$$

Hence equation (1.05) or (1.43) of chapter one becomes,

$$e_t = y_t - x_t - s_t \qquad (t=1,2,\ldots,T) \qquad (6.02)$$

which in matrix terms can be written in a similar way to equations (1.08) and (1.45) of chapter one, i.e.

$$\underline{e} = \underline{y} - \underline{x} - \underline{s} \qquad (6.03)$$

where

$$\begin{aligned} \underline{y}^T &= (y_T,\ldots,y_2,y_1) \\ \underline{x}^T &= (x_T,\ldots,x_2,x_1) \\ \underline{s}^T &= (s_T,\ldots,s_2,s_1) \\ \underline{e}^T &= (e_T,\ldots,e_2,e_1) \end{aligned} \qquad \ldots.(6.04)$$

6.12 SMOOTHNESS

The structural trend or smoothness equation remains in its general autoregressive form as equation (1.42) of chapter one as:

$$x_t = \vartheta_1.x_{t-1} + \vartheta_2.x_{t-2} + \ldots + \vartheta_d.x_{t-d} + a_t \qquad (6.05)$$

where t takes values of d+1 to T.

or in matrix terms as equation (1.46) of chapter one, i.e.

$$\underline{a} = \mathbf{D}.\underline{x} \qquad (6.06)$$

where $\underline{a}^T = (a_T, \ldots . a_{d+1})$, and **D** is a T-d × T matrix with structure:

$$\mathbf{D} = \begin{bmatrix} 1 , -\vartheta_1, -\vartheta_2 \ldots\ldots, -\vartheta_d, 0 , 0 , 0 , \ldots\ldots, 0 \\ 0 , 1 , -\vartheta_1, -\vartheta_2 \ldots\ldots, -\vartheta_d, 0 , 0 , 0 , .., 0 \\ 0 , 0 , \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{etc} \end{bmatrix} \qquad (6.07)$$

6.13 SEASONAL SMOOTHNESS

The equivalent equation to (6.05) for seasonal smoothness or structure can be modelled in several ways, fortunately none of which need be specified for the purposes of this chapter.

For example, it could be modelled as a general autoregressive seasonal formulation, i.e.

$$s_t = \phi_1.s_{t-s} + \phi_2.s_{t-2s} + \ldots + \phi_l.s_{t-ks} + u_t \qquad (6.08)$$

where "s" is the seasonal period, "k" is the seasonal lag and the values of t range from ks+1 to T.

Alternatively, it could be written such that, the seasonal residuals, u_t, for t=s-1 to T, measure the current "yearly" seasonal residual, i.e.

$$u_t = s_t + s_{t-1} + \ldots + s_{t-s+1} \qquad (6.09)$$

To cover all possibilities, it is sufficient at this stage to adopt another general autoregressive model, as was done for trend, (although in practice the trend and seasonal structures would be distinct), namely,

$$s_t = \phi_1.s_{t-1} + \phi_2.s_{t-2} + \ldots + \phi_p.s_{t-p} + u_t \qquad (6.10)$$

where t takes values of p+1 to T.

In matrix terms, we can write this as,

$$\underline{u} = \mathbf{P}.\underline{s} \qquad (6.11)$$

where $\underline{u}^T = (u_T,\ldots,u_{p+1})$, and **P** is an appropriate T-p × T matrix with the same structure as the matrix **D** of (6.07).

6.14 WEIGHTED LEAST SQUARES

Whittaker's approach now requires us to minimise a weighted sum of the fidelity, smoothness and seasonal smoothness sums of squared residuals, i.e. to minimise ψ where,

$$\psi = \underline{e}^T.\underline{e} + (1/\omega).\underline{a}^T.\underline{a} + (1/\upsilon).\underline{u}^T.\underline{u} \qquad (6.12)$$

Using (6.03), (6.06) and (6.11) equation (6.12) can be written,

$$\psi = (\underline{y}-\underline{x}-\underline{s})^T.(\underline{y}-\underline{x}-\underline{s}) + (1/\omega).\underline{x}^T.\mathbf{D}^T.\mathbf{D}.\underline{x} + (1/\upsilon).\underline{s}^T.\mathbf{P}^T.\mathbf{P}.\underline{s} \qquad (6.13)$$

Differentiating ψ in (6.13) with respect to $\underline{x}$ and $\underline{s}$ gives us equations for their respective estimates, $\hat{\underline{x}}$ and $\hat{\underline{s}}$, i.e.

$$2.(\underline{y}-\hat{\underline{x}}-\hat{\underline{s}}) = (2/\omega).\mathbf{D}^T.\mathbf{D}.\hat{\underline{x}} = (2/\upsilon).\mathbf{P}^T.\mathbf{P}.\hat{\underline{s}} \qquad (6.14)$$

Rearranging (6.14) to be consistent with equation (1.37) of chapter one, we get,

$$\underline{y} = \hat{\underline{s}} + \Pi_T.\hat{\underline{x}} = \hat{\underline{x}} + \Upsilon_T.\hat{\underline{s}} \qquad (6.15)$$

where $\Pi_T = \mathbf{I}_T + 1/\omega.\mathbf{D}^T.\mathbf{D}$ and $\Upsilon_T = \mathbf{I}_T + 1/\upsilon.\mathbf{P}^T.\mathbf{P}$, $\mathbf{I}_T$ being a TxT identity matrix.

Rewriting (6.15) in matrix form and inverting we get,

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \Pi_T & \mathbf{I}_T \\ \mathbf{I}_T & \Upsilon_T \end{bmatrix}^{-1} . \begin{bmatrix} \underline{y} \\ \underline{y} \end{bmatrix} \qquad (6.16)$$

From which it follows that,

$$\hat{\underline{x}} = (\Upsilon_T.\Pi_T - \mathbf{I}_T)^{-1}.(\Upsilon_T-\mathbf{I}_T).\underline{y} \qquad (6.17)$$

And similarly,

$$\hat{\underline{s}} = (\Pi_T.\Upsilon_T - \mathbf{I}_T)^{-1}.(\Pi_T-\mathbf{I}_T).\underline{y} \qquad (6.18)$$

which are equivalent to Whittaker's, (weighted least squares), estimates for the seasonal case.

6.2 DISTRIBUTIONAL ASSUMPTIONS

As was the case in chapter one, in considering other estimation approaches, the only assumptions we need make regarding the elements of the residual vectors $\underline{e}$, $\underline{a}$ and $\underline{u}$ of equations (6.03), (6.06) and (6.11) are that they are all independently distributed with zero means and variances σ_e^2, σ_a^2 and σ_u^2 respectively, i.e. we do not need to assume Normality at this stage.

Hence the joint covariance matrix of their combined random vector is given by:

$$\mathbb{C}ov\begin{bmatrix} \underline{E} \\ \underline{A} \\ \underline{U} \end{bmatrix} = \begin{bmatrix} \sigma_e^2.\mathbf{I}_T & \emptyset & \emptyset \\ \emptyset & \sigma_a^2.\mathbf{I}_{T-d} & \emptyset \\ \emptyset & \emptyset & \sigma_u^2.\mathbf{I}_{T-p} \end{bmatrix} \qquad (6.19)$$

6.21 GENERALISED LEAST SQUARES (MMSE)

Equations (6.03), (6.06) and (6.11) can be jointly written in the usual regression form as:

$$\begin{bmatrix} \underline{y} \\ \emptyset \\ \emptyset \end{bmatrix} = \begin{bmatrix} \mathbf{I}_T & \mathbf{I}_T \\ \mathbf{D} & \emptyset \\ \emptyset & \mathbf{P} \end{bmatrix} . \begin{bmatrix} \underline{x} \\ \\ \underline{s} \end{bmatrix} + \begin{bmatrix} \underline{e} \\ -\underline{a} \\ -\underline{u} \end{bmatrix} \qquad (6.20)$$

Applying the results of the appendix to chapter one for linear, unbiased estimators of stochastic parameters, we obtain the following minimum mean square unbiased estimates $\hat{\mathbf{x}}$ and $\hat{\mathbf{s}}$ for $\underline{x}$ and $\underline{s}$,

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{s}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{I}_T & \mathbf{I}_T \\ \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \end{bmatrix}^T . \begin{bmatrix} \sigma_e^2.\mathbf{I}_T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_a^2.\mathbf{I}_{T-d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_u^2.\mathbf{I}_{T-p} \end{bmatrix}^{-1} . \begin{bmatrix} \mathbf{I}_T & \mathbf{I}_T \\ \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \end{bmatrix} \right]^{-1} .\text{continued.}$$

$$\begin{bmatrix} \mathbf{I}_T & \mathbf{I}_T \\ \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \end{bmatrix}^T . \begin{bmatrix} \sigma_e^2.\mathbf{I}_T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_a^2.\mathbf{I}_{T-d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_u^2.\mathbf{I}_{T-p} \end{bmatrix}^{-1} . \begin{bmatrix} \underline{y} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \ldots . (6.21)$$

Simplifying (6.21) gives,

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \Pi_T & \mathbf{I}_T \\ \mathbf{I}_T & \Upsilon_T \end{bmatrix}^{-1} . \begin{bmatrix} \underline{y} \\ \underline{y} \end{bmatrix} \qquad (6.22)$$

if Π_T and Υ_T are defined as in (6.15) but with $\omega=\sigma_a^2/\sigma_e^2$ and $\upsilon=\sigma_u^2/\sigma_e^2$.

Hence comparing (6.22) with (6.16) we see that the Generalised Least Squares, (MMSE), estimates are exactly equivalent to Whittaker's, (weighted least squares), estimates of equations (6.17) and (6.18), i.e.

$$\boxed{\hat{\mathbf{x}} = (\Upsilon_T.\Pi_T - \mathbf{I}_T)^{-1}.(\Upsilon_T - \mathbf{I}_T).\underline{y}} \qquad (6.23)$$

and

$$\boxed{\hat{\mathbf{s}} = (\Pi_T.\Upsilon_T - \mathbf{I}_T)^{-1}.(\Pi_T - \mathbf{I}_T).\underline{y}} \qquad (6.24)$$

In other words the weighted least squared estimates will be minimum mean squared linear estimates if all residuals are assumed to be independent of each other, with zero means and respective variances σ_e^2, σ_a^2 and σ_u^2, with weights, ω and υ of (6.12), chosen such that $\omega=\sigma_a^2/\sigma_e^2$ and $\upsilon=\sigma_u^2/\sigma_e^2$.

The mean squared error matrix of $\hat{\mathbf{x}}$ and $\hat{\mathbf{s}}$ is also given from the results in the appendix to chapter one as:

$$\text{MSE}\begin{bmatrix}\hat{\mathbf{x}}\\ \hat{\mathbf{s}}\end{bmatrix} = \sigma_e^2.\begin{bmatrix}\Pi_T & I_T\\ I_T & \Upsilon_T\end{bmatrix}^{-1} = \sigma_e^2.\begin{bmatrix}\Upsilon_T.(\Pi_T.\Upsilon_T-I_T)^{-1} & -(\Upsilon_T.\Pi_T-I_T)^{-1}\\ -(\Pi_T.\Upsilon_T-I_T)^{-1} & \Pi_T.(\Upsilon_T.\Pi_T-I_T)^{-1}\end{bmatrix} \quad \ldots.(6.25)$$

Note that the inverted matrix in (6.25) is symmetric since,

$$(\Upsilon_T.\Pi_T-I_T)^T = \Pi_T.\Upsilon_T-I_T \qquad (6.26)$$

because Υ_T and Π_T are symmetric from their definitions in (6.15).

Note also the relations,

$$\Pi_T.(\Upsilon_T.\Pi_T-I_T)^{-1} = (\Pi_T.\Upsilon_T-I_T)^{-1}.\Pi_T \qquad (6.27)$$

$$\Upsilon_T.(\Pi_T.\Upsilon_T-I_T)^{-1} = (\Upsilon_T.\Pi_T-I_T)^{-1}.\Upsilon_T \qquad (6.28)$$

since

$$\Pi_T.(\Upsilon_T.\Pi_T-I_T) = (\Pi_T.\Upsilon_T-I_T).\Pi_T$$

$$\Upsilon_T.(\Pi_T.\Upsilon_T-I_T) = (\Upsilon_T.\Pi_T-I_T).\Upsilon_T$$

Finally note that since the matrix, $(\Pi_T.\Upsilon_T-I_T)$, is non-zero in (6.25), the estimates $\hat{\mathbf{x}}$ and $\hat{\mathbf{s}}$ are correlated even though their true values $\mathbf{x}$ and $\mathbf{s}$ are, from their generated mechanisms of (6.06) and (6.11), obviously independent. This should not be surprising since the regression model always produces estimated residuals which are correlated.

6.22 STANDARDISED RESIDUALS

Writing (6.12) with ω and υ defined by $\omega=\sigma_a^2/\sigma_e^2$ and $\upsilon=\sigma_u^2/\sigma_e^2$ gives,

$$\psi = \underline{e}^T.\underline{e} + \sigma_e^2/\sigma_a^2.\underline{a}^T.\underline{a} + \sigma_e^2/\sigma_u^2.\underline{u}^T.\underline{u}$$

$$\text{i.e.} \qquad \psi = \sigma_e^2.(\underline{e}^T.\underline{e}/\sigma_e^2 + \underline{a}^T.\underline{a}/\sigma_a^2 + \underline{u}^T.\underline{u}/\sigma_u^2) = \sigma_e^2.\psi^*$$

$$\text{where} \qquad \psi^* = \underline{e}^T.\underline{e}/\sigma_e^2 + \underline{a}^T.\underline{a}/\sigma_a^2 + \underline{u}^T.\underline{u}/\sigma_u^2 \qquad (6.29)$$

Since minimisation of ψ with respect to $\underline{x}$ and $\underline{s}$ is exactly equivalent to minimisation of ψ^*, we see from (6.29) that minimising Whittaker's function is equivalent to minimising the sum of squares of <u>standardised</u> residuals.

6.3 THE EQUIVALENCE OF CLASSICAL, (GLS), AND STATE SPACE ESTIMATION

In chapters two and three we saw that State Space estimates were simply the means of the posterior distributions of the parameters i.e. for each of the trend values x_t, t=1 to T, the estimates were given by the means of the conditional distributions of x_t given all observed values y_1 to y_T, which we wrote concisely as the distribution of $x_t(T)$. To do this we needed to specify the prior, (i.e. given no observed values of y_t), distribution of the initial values of x_t which would be required in order to generate the later values, i.e. $x_1(0)$, $x_2(0)$, ... , $x_d(0)$, where "d" is the degree of differencing defined in equation (6.04). The ideas follow much the same lines as those of section **3.1** of chapter three.

Extending the estimation procedure to include seasonality requires us to further specify the prior distribution of the initial "p" values of the seasonal component, $s_1(0)$, $s_2(0)$, ... , $s_p(0)$, where the value of "p" is explained in section **6.13** and depends on the actual seasonal form chosen.

6.31 THE DISTRIBUTION OF $\mathbf{x}_T(0)$ given $\mathbf{x}_d(0)$

Suppose that the the prior distributions of the vectors $\mathbf{x}_d(0)$ and $\mathbf{s}_p(0)$ have means μ_d and μ_p, and covariance matrices Σ_d and Σ_p where $\mathbf{x}_d(0)$ and $\mathbf{s}_p(0)$ have the following elements,

$$\mathbf{x}_d(0) = (x_d(0),\ x_{d-1}(0),\ldots,x_1(0))^T \qquad (6.30)$$

$$\mathbf{s}_p(0) = (s_p(0),\ s_{p-1}(0),\ldots,s_1(0))^T \qquad (6.31)$$

Then using (6.06), and assuming the residuals a_t are independently generated, we may write, (in an similar way to equations (3.04) to (3.07) of chapter three),

$$\mathbf{a}_{T-d}(0) = \mathbf{D}.\mathbf{x}_T(0) = \mathbf{B}.\mathbf{x}_{T-d}(0) - \mathbf{C}.\mathbf{x}_d(0) \qquad (6.32)$$

where,

$$\begin{aligned} \mathbf{a}_{T-d}(0) &= (a_T(0),\ \ldots\ ,\ a_d(0))^T \\ \mathbf{x}_T(0) &= (x_T(0),\ \ldots\ ,\ x_1(0))^T \\ \mathbf{x}_{T-d}(0) &= (x_T(0),\ \ldots\ ,\ x_d(0))^T \end{aligned} \qquad \ldots.(6.33)$$

and the matrix $\mathbf{D}$ has been partitioned into two matrices $\mathbf{B}$ and $\mathbf{C}$, i.e. $\mathbf{D} = [\mathbf{B}\vdots-\mathbf{C}]$, where $\mathbf{B}$ is now a square T-d $\times$ T-d invertible matrix, as follows:

$$\mathbf{D} = [\mathbf{B}\vdots-\mathbf{C}] = \left[\begin{array}{c:c} 1,-\vartheta_1,\ldots-\vartheta_d,\ 0\ ,\ldots\ldots,\ 0 & 0\ ,\ \ldots\ 0,\ 0 \\ & \\ 0,\ldots\ldots\ldots\ldots 0\ ,\ 1\ ,-\vartheta_1 & -\vartheta_2,\ldots-\vartheta_{d-1},0 \\ 0,\ 0\ ,\ldots\ldots\ \ \ldots\ .0\ ,\ 1 & -\vartheta_1,\ \ \ldots\ ,-\vartheta \end{array}\right] \qquad (6.34)$$

Hence, since $\mathbf{B}$ is invertible, we can rearrange (6.32) to give,

$$\mathbf{x}_{T-d}(0) = \mathbf{B}^{-1}.\mathbf{C}.\mathbf{x}_d(0) + \mathbf{B}^{-1}.\mathbf{a}_{T-d}(0) \qquad (6.35)$$

Taking expectations of (6.35) and remembering that $\mathbb{E}[\mathbf{x}_d(0)] = \mu_d$, gives,

$$\mathbb{E}[\mathbf{x}_{T-d}(0)] = \mathbf{B}^{-1}.\mathbf{C}.\mu_d \tag{6.36}$$

Hence,

$$\mathbb{E}[\mathbf{x}_T(0)] = \mu_{\mathbf{x}}(0) = \mathbb{E}\begin{bmatrix} \mathbf{x}_{T-d}(0) \\ \mathbf{x}_d(0) \end{bmatrix} = \begin{bmatrix} \mathbf{B}^{-1}.\mathbf{C}.\mu_d \\ \mu_d \end{bmatrix} \tag{6.37}$$

Since the elements of $\mathbf{x}_d(0)$ and $\mathbf{a}_{T-d}(0)$ are independent, because all values of $\mathbf{a}_{T-d}(0)$ are generated after $\mathbf{x}_d(0)$, then,

$$\mathbb{C}ov[\mathbf{x}_{T-d}(0)] = \mathbf{B}^{-1}.\mathbf{C}.\Sigma_d.\mathbf{C}^T.(\mathbf{B}^T)^{-1} + \sigma_a^2.\mathbf{B}^{-1}.(\mathbf{B}^T)^{-1} \tag{6.38}$$

$$\mathbb{C}ov[\mathbf{x}_{T-d}(0)] = \mathbf{B}^{-1}.(\mathbf{C}.\Sigma_d.\mathbf{C}^T + \sigma_a^2.\mathbf{I}_{T-d}).(\mathbf{B}^T)^{-1}$$

$$\mathbb{C}ov[\mathbf{x}_{T-d}(0).\mathbf{x}_d(0)^T] = \mathbf{B}^{-1}.\mathbf{C}.\Sigma_d \tag{6.39}$$

Hence,

$$\Sigma_{\mathbf{x}}(0) = \mathbb{C}ov[\mathbf{x}_T(0)] = \mathbb{C}ov\begin{bmatrix} \mathbf{x}_{T-d}(0) \\ \mathbf{x}_d(0) \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{B}^{-1}.(\mathbf{C}.\Sigma_d.\mathbf{C}^T + \sigma_a^2.\mathbf{I}_{T-d}).(\mathbf{B}^T)^{-1} & \mathbf{B}^{-1}.\mathbf{C}.\Sigma_d \\ \Sigma_d.\mathbf{C}^T.(\mathbf{B}^T)^{-1} & \Sigma_d \end{bmatrix}$$

$$= \begin{bmatrix} 1/\sigma_a^2.\mathbf{B}^T.\mathbf{B} & -1/\sigma_a^2.\mathbf{B}^T.\mathbf{C} \\ -1/\sigma_a^2.\mathbf{C}^T.\mathbf{B} & 1/\sigma_a^2.\mathbf{C}^T.\mathbf{C} + \Sigma_d^{-1} \end{bmatrix}^{-1} \tag{6.40}$$

If we now assume that $\mathbf{x}_d(0)$ has a vague prior distribution, i.e. we let the variances of each of its individual elements tend to infinity, then this implies that Σ_d^{-1} will tend to zero, and from (6.40) and the definition of **D** in (6.34) we see that,

$$\Sigma_{\mathbf{x}}(0)^{-1} \rightarrow \begin{bmatrix} 1/\sigma_a^2.\mathbf{B}^T.\mathbf{B} & -1/\sigma_a^2.\mathbf{B}^T.\mathbf{C} \\ -1/\sigma_a^2.\mathbf{C}^T.\mathbf{B} & 1/\sigma_a^2.\mathbf{C}^T.\mathbf{C} \end{bmatrix} = 1/\sigma_a^2.\mathbf{D}^T.\mathbf{D} \qquad (6.41)$$

In summary, therefore, the vector $\mathbf{x}_T(0)$, given a vague prior for $\mathbf{x}_d(0)$, has mean $\mu_{\mathbf{x}}(0)$ given by (6.37) and covariance matrix $\Sigma_{\mathbf{x}}(0)$, given by (6.41).

6.32 THE DISTRIBUTION OF $\mathbf{s}_T(0)$ given $\mathbf{s}_p(0)$

A similar procedure to the above can be followed to produce the mean, $\mu_{\mathbf{s}}(0)$, and covariance matrix, $\Sigma_{\mathbf{s}}(0)$, of $\mathbf{s}_T(0)$, given a vague prior for $\mathbf{s}_T(0)$, where $\mathbf{s}_T(0) = (s_T(0), \ldots, s_1(0))^T$.

The parallel result to (6.41) is given by,

$$\mathbb{C}ov[\mathbf{s}_T(0)]^{-1} = \Sigma_{\mathbf{s}}(0)^{-1} \rightarrow 1/\sigma_u^2.\mathbf{P}^T.\mathbf{P} \text{ as } \Sigma_p(0)^{-1} \rightarrow \varnothing \qquad (6.42)$$

Also, by partitioning the T-p × T matrix **P** so that $\mathbf{P} = [\mathbf{E}\,|\,-\mathbf{F}]$, where **E** is a square T-p × T-p invertible matrix, we can produce the equivalent result to (6.37), i.e.

$$\mathbb{E}[\mathbf{s}_T(0)] = \mu_{\mathbf{s}}(0) = \mathbb{E}\begin{bmatrix} \mathbf{s}_{T-p}(0) \\ \mathbf{s}_p(0) \end{bmatrix} = \begin{bmatrix} \mathbf{E}^{-1}.\mathbf{F}.\mu_p \\ \mu_p \end{bmatrix} \qquad (6.43)$$

6.33 BEST LINEARLY CONDITIONAL ESTIMATES

Conditioning (6.03) on zero values of y_t, we can write

$$\mathbf{y}_T(0) = \mathbf{x}_T(0) + \mathbf{s}_T(0) + \mathbf{e}_T(0) \qquad (6.44)$$

where $\underline{y}$, $\underline{x}$, $\underline{s}$ and $\underline{e}$ are alternatively written $\mathbf{y}_T$, $\mathbf{x}_T$, $\mathbf{s}_T$ and $\mathbf{e}_T$.

From (6.44) we have, on taking expectations, and using (6.37) and (6.43),

$$\mathbb{E}[\mathbf{y}_T(0)] = \mathbb{E}[\mathbf{x}_T(0)] + \mathbb{E}[\mathbf{s}_T(0)] = \mu_{\mathbf{x}}(0) + \mu_{\mathbf{s}}(0) \qquad (6.45)$$

Also from (6.44), taking into account that $\mathbf{x}_T(0)$, $\mathbf{s}_T(0)$ and $\mathbf{e}_T(0)$ are all independent, and using (6.41) and (6.42),

$$\mathbb{C}ov[\mathbf{y}_T(0)] = \mathbb{C}ov[\mathbf{x}_T(0)] + \mathbb{C}ov[\mathbf{s}_T(0)] + \sigma_e^2.\mathbf{I}_T$$

$$= \Sigma_{\mathbf{x}}(0) + \Sigma_{\mathbf{s}}(0) + \sigma_e^2.\mathbf{I}_T \qquad (6.46)$$

$$\mathbb{C}ov[\mathbf{y}_T(0).\mathbf{x}_T(0)^T] = \mathbb{C}ov[\mathbf{x}_T(0)] = \Sigma_{\mathbf{x}}(0) \qquad (6.47)$$

$$\mathbb{C}ov[\mathbf{y}_T(0).\mathbf{s}_T(0)^T] = \mathbb{C}ov[\mathbf{s}_T(0)] = \Sigma_{\mathbf{s}}(0) \qquad (6.48)$$

Hence using results (6.45) to (6.48), we obtain the joint distribution of $\mathbf{y}_T(0)$, $\mathbf{x}_T(0)$ and $\mathbf{s}_T(0)$ as,

$$\begin{bmatrix} \mathbf{y}_T(0) \\ \mathbf{x}_T(0) \\ \mathbf{s}_T(0) \end{bmatrix} \sim UD \left[\begin{array}{ccc|ccc} \mu_{\mathbf{x}}(0) + \mu_{\mathbf{s}}(0) & ; & \Sigma_{\mathbf{x}}(0)+\Sigma_{\mathbf{s}}(0)+\sigma_e^2.\mathbf{I}_T & , & \Sigma_{\mathbf{x}}(0) & , & \Sigma_{\mathbf{s}}(0) \\ \mu_{\mathbf{x}}(0) & ; & \Sigma_{\mathbf{x}}(0) & , & \Sigma_{\mathbf{x}}(0) & , & \varnothing \\ \mu_{\mathbf{s}}(0) & ; & \Sigma_{\mathbf{s}}(0) & , & \varnothing & , & \Sigma_{\mathbf{s}}(0) \end{array} \right]$$

....(6.49)

We can now apply the results of section **2.4** of chapter two to obtain the "best" linearly joint conditional distribution of $\mathbf{x}_T(T)$ and $\mathbf{s}_T(T)$, i.e. the joint distribution of $\mathbf{x}_T$ and $\mathbf{s}_T$ based on all T values of $\mathbf{y}_T$.

The covariance matrix of $\mathbf{x}_T(T)$ and $\mathbf{s}_T(T)$ is given by:

$$\mathbb{C}ov\begin{bmatrix} \mathbf{x}_T(T) \\ \mathbf{s}_T(T) \end{bmatrix} = \begin{bmatrix} \Sigma_{\mathbf{x}}(T) & , & \Sigma_{\mathbf{xs}}(T) \\ \Sigma_{\mathbf{sx}}(T) & , & \Sigma_{\mathbf{s}}(T) \end{bmatrix}$$

where,

$$\Sigma_{\mathbf{x}}(T) = \Sigma_{\mathbf{x}}(0) - \Sigma_{\mathbf{x}}(0).(\Sigma_{\mathbf{x}}(0) + \Sigma_{\mathbf{s}}(0) + \sigma_e^2.I_T)^{-1}.\Sigma_{\mathbf{x}}(0)$$

$$\Sigma_{\mathbf{s}}(T) = \Sigma_{\mathbf{s}}(0) - \Sigma_{\mathbf{s}}(0).(\Sigma_{\mathbf{x}}(0) + \Sigma_{\mathbf{s}}(0) + \sigma_e^2.I_T)^{-1}.\Sigma_{\mathbf{s}}(0)$$

$$\Sigma_{\mathbf{xs}}(T) = - \Sigma_{\mathbf{x}}(0).(\Sigma_{\mathbf{x}}(0) + \Sigma_{\mathbf{s}}(0) + \sigma_e^2.I_T)^{-1}.\Sigma_{\mathbf{s}}(0)$$

$$\Sigma_{\mathbf{sx}}(T) = - \Sigma_{\mathbf{s}}(0).(\Sigma_{\mathbf{x}}(0) + \Sigma_{\mathbf{s}}(0) + \sigma_e^2.I_T)^{-1}.\Sigma_{\mathbf{x}}(0)$$

....(6.50)

Defining Π_T and Υ_T as in (6.15) with $\omega=\sigma_a^2/\sigma_e^2$ and $\upsilon=\sigma_u^2/\sigma_e^2$. and using (6.41) and (6.42) we get the following relations,

$$\Sigma_{\mathbf{x}}(0)^{-1} = (\Pi_T - I_T)/\sigma_e^2 \text{ and } \Sigma_{\mathbf{s}}(0)^{-1} = (\Upsilon_T - I_T)/\sigma_e^2 \qquad (6.51)$$

Applying (6.51) to (6.50) we arrive, after some messy algebra, at

$$\Sigma_{\mathbf{x}}(T) = \sigma_e^2.\Upsilon_T.(\Pi_T.\Upsilon_T - I_T)^{-1} \qquad \Sigma_{\mathbf{xs}}(T) = - \sigma_e^2.(\Upsilon_T.\Pi_T - I_T)^{-1}$$

$$\Sigma_{\mathbf{sx}}(T) = - \sigma_e^2.(\Pi_T.\Upsilon_T - I_T)^{-1} \qquad \Sigma_{\mathbf{s}}(T) = \sigma_e^2.\Pi_T.(\Upsilon_T.\Pi_T - I_T)^{-1}$$

....(6.52)

Hence, the posterior covariances in (6.52) are therefore exactly equal to the mean squared errors of (6.25).

Section **2.4** of chapter two also gives us the posterior means of $\mathbf{x}_T(T)$ and $\mathbf{s}_T(T)$ as:

$$\mathbb{E}\begin{bmatrix} \mathbf{x}_T(T) \\ \mathbf{s}_T(T) \end{bmatrix} = \begin{bmatrix} \mu_{\mathbf{x}}(T) \\ \mu_{\mathbf{s}}(T) \end{bmatrix}$$

where,

$$\mu_x(T) = \mu_x(0) - \Sigma_x(0).(\Sigma_x(0)+\Sigma_s(0)+\sigma_e^2.I_T)^{-1}.(y_T-\mu_x(0)-\mu_s(0))$$

$$\mu_s(T) = \mu_s(0) - \Sigma_s(0).(\Sigma_x(0)+\Sigma_s(0)+\sigma_e^2.I_T)^{-1}.(y_T-\mu_x(0)-\mu_s(0))$$

....(6.53)

Applying the results of (6.50) to (6.53) we get,

$$\mu_x(T) = \Sigma_x(T).\Sigma_x(0)^{-1}.\mu_x(0) + \Sigma_{xs}(T).\Sigma_s(0)^{-1}.\mu_s(0) - \Sigma_{xs}(T).\Sigma_s(0)^{-1}.y_T$$

$$\mu_s(T) = \Sigma_s(T).\Sigma_s(0)^{-1}.\mu_s(0) + \Sigma_{sx}(T).\Sigma_x(0)^{-1}.\mu_x(0) - \Sigma_{sx}(T).\Sigma_x(0)^{-1}.y_T$$

....(6.54)

However from (6.34) and (6.37) we find,

$$D.\mu_x(0) = [B\vdots-C].\begin{bmatrix} B^{-1}.C.\mu_d \\ \mu_d \end{bmatrix} = C.\mu_d - C.\mu_d = \varnothing \qquad (6.55)$$

and so using (6.41),

$$\Sigma_x(0)^{-1}.\mu_x(0) = 1/\sigma_a^2.D^T.D.\mu_x(0) = \varnothing \qquad (6.56)$$

Similarly we can show that,

$$P.\mu_s(0) = \Sigma_s(0)^{-1}.\mu_s(0) = \varnothing \qquad (6.57)$$

Applying (6.56) and (6.57) to (6.54) we find,

$$\mu_x(T) = -\Sigma_{xs}(T).\Sigma_s(0)^{-1}.y_T$$

$$\mu_s(T) = -\Sigma_{sx}(T).\Sigma_x(0)^{-1}.y_T \qquad (6.58)$$

which finally give, using (6.51) and (6.52)

$$\mu_{\mathbf{x}}(T) = (\Upsilon_T.\Pi_T - I_T)^{-1}.(\Upsilon_T - I_T).\mathbf{y}_T$$

$$\mu_{\mathbf{s}}(T) = (\Upsilon_T.\Pi_T - I_T)^{-1}.(\Pi_T - I_T).\mathbf{y}_T$$

....(6.59)

Equations (6.59) are exactly the same as those obtained for the minimum mean squared unbiased estimates in (6.23) and (6.24).

The overall conclusion is that Classical, (MMSE), and State Space estimation procedures will produce exactly the same results as long as the State Space approach assumes that initial trend and seasonal components have vague prior distributions i.e. they have infinite variances. Note that their initial means are of no consequence as long as they are finite.

This completes the extension of seasonality to the main results of chapters one to three. Note that in doing so we have not had to assume Normality, only linearity. As was also explained in those chapters, if Normality can be assumed, the same results will be obtained but without the need to assume linearity.

The next sections go on to extend the seasonal case to the results of chapters four and five, i.e. the estimation of the residual variances, where Normality is an essential pre-requisite.

6.4 THE ESTIMATION OF RESIDUAL VARIANCES

6.41 THE RELATIONSHIP BETWEEN DATA AND RESIDUALS

As was the case in chapter four, we firstly need to establish a relationship between the data vector $\mathbf{y}_T$ and the vectors of residuals $\mathbf{e}_T$, $\mathbf{a}_{T-d}$ and $\mathbf{u}_{T-p}$ where:

$$\begin{aligned} \mathbf{y}_T &= (y_T,\ y_{T-1}\ \ldots\ ,\ y_1)^T \\ \mathbf{e}_T &= (e_T,\ e_{T-1}\ \ldots\ ,\ e_1)^T \\ \mathbf{a}_{T-d} &= (a_T,\ a_{T-1}\ \ldots\ ,\ a_{d+1})^T \\ \mathbf{u}_{T-p} &= (u_T,\ u_{T-1}\ \ldots\ ,\ u_{p+1})^T \end{aligned}$$

....(6.60)

The smoothness relationship of (6.05) can be written as,

$$\Theta_d(\mathbb{B}).x_t = (1 - \vartheta_1.\mathbb{B} - \vartheta_2.\mathbb{B}^2 - \ldots - \vartheta_d.\mathbb{B}^d).x_t = a_t,\ t= d+1 \text{ to } T$$

where $\mathbb{B}$ is the backward operator such that $\mathbb{B}^k.x_t = x_{t-k}$ (6.61)

Similarly the seasonal smoothness relationship in (6.10) can be written,

$$\Phi_p(\mathbb{B}).s_t = (1 - \phi_1.\mathbb{B} - \phi_2.\mathbb{B}^2 - \ldots - \phi_p.\mathbb{B}^d).s_t = u_t,\ t = p+1 \text{ to } T$$

....(6.62)

Because of the multiplicative natures of the operator functions $\Theta_d(\mathbb{B})$ and $\Phi_p(\mathbb{B})$ in (6.61) and (6.62), we can apply them to the measurement equation of (6.01) to give,

$$\Theta_d(\mathbb{B}).\Phi_p(\mathbb{B}).y_t = \Theta_d(\mathbb{B}).\Phi_p(\mathbb{B}).x_t + \Theta_d(\mathbb{B}).\Phi_p(\mathbb{B}).s_t + \Theta_d(\mathbb{B}).\Phi_p(\mathbb{B}).e_t$$

$$\text{for } t = d+p+1 \text{ to } T \qquad (6.63)$$

which since $\Theta_d(\mathbb{B}).\Phi_p(\mathbb{B}) = \Phi_p(\mathbb{B}).\Theta_d(\mathbb{B}) = \Xi_{d+p}(\mathbb{B})$, can be written,

$$\Xi_{d+p}(\mathbb{B}).y_t = \Phi_p(\mathbb{B}).\Theta_d(\mathbb{B}).x_t + \Theta_d(\mathbb{B}).\Phi_p(\mathbb{B}).s_t + \Xi_{d+p}(\mathbb{B}).e_t \qquad (6.64)$$

Substituting (6.61) and (6.62) into (6.64) we get,

$$\Xi_{d+p}(\mathbb{B}).y_t = \Phi_p(\mathbb{B}).a_t + \Theta_d(\mathbb{B}).u_t + \Xi_{d+p}(\mathbb{B}).e_t$$

$$\text{for } t = d+p+1 \text{ to } T \qquad (6.65)$$

Note that the series $\Xi_{d+p}(\mathbb{B}).y_t$ is stationary, (even though y_t itself is not), since it is able to be defined in terms of the sum of, what are essentially, three finite autoregressive white noise processes.

In matrix terms (6.65) can be written in terms of the vector of (6.60) as,

$$\mathbb{G}.\mathbf{y}_T = \mathbb{P}.\mathbf{a}_{T-d} + \mathbb{D}.\mathbf{u}_{T-p} + \mathbb{G}.\mathbf{e}_T \qquad (6.66)$$

where $\mathbb{G}$ is a T-d-p $\times$ T matrix such that,

$$\mathbb{G} = \mathbb{P}.\mathbf{D} = \mathbb{D}.\mathbf{P} \qquad (6.67)$$

$\mathbb{D}$ is a T-d-p $\times$ T-p matrix and $\mathbf{D}$ is the T-d $\times$ T matrix defined in (6.07), both of which have the same structure, i.e

$$\mathbb{D},\ \mathbf{D} = \begin{bmatrix} 1 & -\vartheta_1 & -\vartheta_2 & \ldots\ldots & -\vartheta_d & 0 & 0 & 0 & \ldots\ldots & 0 \\ 0 & 1 & -\vartheta_1 & -\vartheta_2 & \ldots\ldots & -\vartheta_d & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & & & \text{etc} \end{bmatrix} \qquad (6.68)$$

and $\mathbb{P}$ is a T-d-p $\times$ T-d matrix and $\mathbf{P}$ is the T-p $\times$ T matrix defined in (6.11), both of which have the same structure of (6.68), but with ϕ_i substituted for ϑ_i.

Given the assumptions that the elements of $\mathbf{a}_{T-d}$, $\mathbf{u}_{T-p}$ and $\mathbf{e}_T$ are independently distributed with zero means and variances σ_a^2, σ_u^2 and σ_e^2 respectively, the vector $\mathbb{G}.\mathbf{y}_T$, (with T-d-p elements), in (6.66) will have a mean vector, ø, i.e. of zeros, and a covariance matrix given by Σ_{Gy} where,

$$\Sigma_{Gy} = \mathbb{C}ov[\mathbb{G}.\mathbf{y}_T] = \sigma_a^2.\mathbb{P}.\mathbb{P}^T + \sigma_u^2.\mathbb{D}.\mathbb{D}^T + \sigma_e^2.\mathbb{G}.\mathbb{G}^T \qquad (6.69)$$

6.42 MAXIMUM LIKELIHOOD ESTIMATION

The log-likelihood of the residual variances σ_a^2, σ_u^2 and σ_e^2 and also the parameter vectors ϑ and ϕ, where,

$$\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_d)^T \text{ and } \phi = (\phi_1, \phi_2, \ldots, \phi_p)^T \qquad (6.70)$$

is produced in the same way as equation (4.12) of chapter four as $\mathbb{LL}[\sigma_a^2,\sigma_u^2,\sigma_e^2,\vartheta,\phi]$, where,

$$\mathbb{LL}[\sigma_a^2,\sigma_u^2,\sigma_e^2,\vartheta,\phi] = -1/2.((T-d-p).\ln(2\pi) + \ln|\Sigma_{Gy}| + \mathbf{y}_T^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T)$$

....(6.71)

Proceeding on the same lines as section **4.4** of chapter four and differentiating (6.71) with respect to σ^2, which represents any of the three residual variances, and setting to zero gives,

$$\mathbb{TR}[\Sigma_{Gy}^{-1}.(\partial\Sigma_{Gy}/\partial\sigma^2)] = \mathbf{y}_T^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.(\partial\Sigma_{Gy}/\partial\sigma^2).\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \qquad (6.72)$$

Using (6.69) we have,

$$\partial\Sigma_{Gy}/\partial\sigma_a^2 = \mathbb{P}.\mathbb{P}^T, \quad \partial\Sigma_{Gy}/\partial\sigma_u^2 = \mathbb{D}.\mathbb{D}^T, \quad \partial\Sigma_{Gy}/\partial\sigma_e^2 = \mathbb{G}.\mathbb{G}^T \qquad (6.73)$$

Hence combining (6.72) and (6.73) we get the three equations,

$$\mathbb{TR}[\Sigma_{Gy}^{-1}.\mathbb{P}.\mathbb{P}^T] = \mathbf{y}_T^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{P}.\mathbb{P}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \qquad (6.74)$$

$$\mathbb{TR}[\Sigma_{Gy}^{-1}.\mathbb{D}.\mathbb{D}^T] = \mathbf{y}_T^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{D}.\mathbb{D}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \qquad (6.75)$$

$$\mathbb{TR}[\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbb{G}^T] = \mathbf{y}_T^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \qquad (6.76)$$

6.43 MINIMUM VARIANCE ESTIMATION

6.431 The Quadratic Form

This section proceeds along the same lines as section **4.5** of chapter four. We seek an estimator $\hat{\sigma}^2$ of the variance σ^2, where,

$$\hat{\sigma}^2 = \mathbf{y}_T^T.\mathbb{G}^T.\mathbf{M}.\mathbb{G}.\mathbf{y}_T \tag{6.77}$$

where **M** is an unspecified square matrix.

Using the results of appendix B in conjunction with equation (6.69), the mean and variance of $\hat{\sigma}^2$ are given by:

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{TR}[\mathbf{M}.\Sigma_{Gy}] \text{ and } \mathbb{V}[\hat{\sigma}^2] = 2.\mathbb{TR}[\mathbf{M}.\Sigma_{Gy}.\mathbf{M}.\Sigma_{Gy}] \tag{6.78}$$

6.432 Conditionally Unbiased Estimation

Minimising $\mathbb{V}[\hat{\sigma}^2]$ subject to the unbiasedness constraint $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$, is equivalent to minimising,

$$\mathbb{V}[\hat{\sigma}^2] - 4.\lambda.(\mathbb{E}[\hat{\sigma}^2] - \sigma^2) \tag{6.79}$$

where $4.\lambda$ is a Lagrangian multiplier.

Substituting the results of (6.78) into (6.79) we get,

$$2.\mathbb{TR}[\mathbf{M}.\Sigma_{Gy}.\mathbf{M}.\Sigma_{Gy}] - 4.\lambda.(\mathbb{TR}[\mathbf{M}.\Sigma_{Gy}] - \sigma^2) \tag{6.80}$$

Differentiating (6.80) by **M** by employing the results of appendix C, setting to zero, and realising that **M** must be symmetric, gives,

$$\Sigma_{Gy}\mathbf{M}.\Sigma_{Gy} = \lambda.\Sigma_{Gy} \text{ i.e. } \mathbf{M}.\Sigma_{Gy} = \lambda.\mathbf{I}_{T-d-p} \tag{6.81}$$

where I_{T-d-p} is a T-d-p $\times$ T-d-p identity matrix.

Differentiation of (6.80) by λ and substituting (6.81) we have,

$$\sigma^2 = \mathbb{TR}[\mathbf{M}.\Sigma_{Gy}] = \lambda.(T-d-p) \tag{6.82}$$

which from (6.81) gives the following expression for $\mathbf{M}$,

$$\mathbf{M} = \sigma^2.\Sigma_{Gy}^{-1}/(T-d-p) \tag{6.83}$$

Finally using (6.77) we obtain,

$$\hat{\sigma}^2 = \sigma^2.\mathbf{y}_T^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T/(T-d-p) \tag{6.84}$$

Utilising (6.69), we can define the matrix Ω_{Gy} as,

$$\Omega_{Gy} = \Sigma_{Gy}/\sigma_e^2 = \mathbb{G}.\mathbb{G}^T + \omega.\mathbb{P}.\mathbb{P}^T + \upsilon.\mathbb{D}.\mathbb{D}^T \tag{6.85}$$

where $\omega = \sigma_a^2/\sigma_e^2$ and $\upsilon = \sigma_u^2/\sigma_e^2$ are residual variance ratios.

Substituting (6.85) into (6.84), we obtain expressions for conditionally unbiased estimates of all three residual variances in terms of the their variance ratios ω and υ, i.e.

$$\hat{\sigma}_e^2 = \mathbf{y}_T^T.\mathbb{G}^T.\Omega_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T/(T-d-p) \tag{6.86}$$

$$\hat{\sigma}_a^2 = \omega.\hat{\sigma}_e^2 = \omega.\mathbf{y}_T^T.\mathbb{G}^T.\Omega_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T/(T-d-p) \tag{6.87}$$

$$\hat{\sigma}_u^2 = \upsilon.\hat{\sigma}_e^2 = \upsilon.\mathbf{y}_T^T.\mathbb{G}^T.\Omega_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T/(T-d-p) \tag{6.88}$$

We can now note the similarities between equations (6.86) to (6.88) for the seasonal case, and equations (4.31) and (4.32) in chapter four for the non-seasonal case.

6.433 Unconditionally Unbiased Estimation

Again we seek an estimator $\hat{\sigma}^2$ of the variance σ^2, of the form described in section 6.44. However this time σ^2 is defined as a weighted sum of the three residual variances, i.e.

$$\sigma^2 = \alpha_a.\sigma_a^2 + \alpha_u.\sigma_u^2 + \alpha_e.\sigma_e^2 \qquad (6.89)$$

the purpose being that by setting two of the α_i to zero and the other to unity we will find an estimate of the particular variance in question.

Using (6.78) in conjunction with (6.69) and appendix B, we have,

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{TR}[\mathbf{M}.\Sigma_{Gy}] = \sigma_a^2.\mathbb{TR}[\mathbf{M}.\mathbb{P}.\mathbb{P}^T] + \sigma_u^2.\mathbb{TR}[\mathbf{M}.\mathbb{D}.\mathbb{D}^T] + \sigma_e^2.\mathbb{TR}[\mathbf{M}.\mathbb{G}.\mathbb{G}^T]$$

....(6.90)

Since the expressions $\mathbf{M}.\mathbb{P}.\mathbb{P}^T$, $\mathbf{M}.\mathbb{D}.\mathbb{D}^T$ and $\mathbf{M}.\mathbb{G}.\mathbb{G}^T$ in (6.90) are independent of the values of σ_a^2, σ_u^2 and σ_e^2, we can ensure that the estimator, $\hat{\sigma}^2$, is unconditionally unbiased by including the conditions that the elements of the vector $\underline{\alpha} = (\alpha_a, \alpha_u, \alpha_e)^T$ in (6.89) are such that,

$$\underline{\alpha} = \begin{pmatrix} \alpha_a \\ \alpha_u \\ \alpha_e \end{pmatrix} = \begin{pmatrix} \mathbb{TR}[\mathbf{M}.\mathbb{P}.\mathbb{P}^T] \\ \mathbb{TR}[\mathbf{M}.\mathbb{D}.\mathbb{D}^T] \\ \mathbb{TR}[\mathbf{M}.\mathbb{G}.\mathbb{G}^T] \end{pmatrix} = \underline{\mathbb{T}}_M \qquad (6.91)$$

Minimising $\mathbb{V}[\hat{\sigma}^2]$ subject to the unbiasedness constraint $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$, is now, using (6.78) and (6.91), equivalent to minimising,

$$2.\mathbb{TR}[\mathbf{M}.\Sigma_{Gy}.\mathbf{M}.\Sigma_{Gy}] - 4.\underline{\lambda}^T.(\underline{\mathbb{T}}_M - \underline{\alpha}) \tag{6.92}$$

where $\underline{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ is a vector of Lagrangian multipliers.

Differentiating (6.92) with respect to **M**, using appendix C, and setting to zero gives,

$$\Sigma_{Gy}.\mathbf{M}^T.\Sigma_{Gy} = \lambda_1.\mathbb{P}.\mathbb{P}^T + \lambda_2.\mathbb{D}.\mathbb{D}^T + \lambda_3.\mathbb{G}.\mathbb{G}^T \tag{6.93}$$

From (6.93) it follows that,

$$\mathbf{M}^T = \lambda_1.\mathbf{M}_P + \lambda_2.\mathbf{M}_D + \lambda_3.\mathbf{M}_G = \underline{\lambda}^T.\mathbf{M}_{PDG} \tag{6.94}$$

where the element matrices of $\mathbf{M}_{PDG} = (\mathbf{M}_P, \mathbf{M}_D, \mathbf{M}_G)^T$ are given by:

$$\mathbf{M}_X = \Sigma_{Gy}^{-1}.\mathbb{X}.\mathbb{X}^T.\Sigma_{Gy}^{-1} \tag{6.95}$$

and therefore **M**, $(=\mathbf{M}^T)$, is symmetric

Post-multiplying (6.94) by $\mathbb{P}.\mathbb{P}^T$, $\mathbb{D}.\mathbb{D}^T$ and $\mathbb{G}.\mathbb{G}^T$ and then taking the traces gives us the elements of $\underline{\mathbb{T}}_M$ and hence $\underline{\alpha}$ in (6.91), i.e.

$$\begin{pmatrix} \alpha_a \\ \alpha_u \\ \alpha_e \end{pmatrix} = \begin{pmatrix} \mathbb{TR}[\mathbf{M}.\mathbb{P}.\mathbb{P}^T] \\ \mathbb{TR}[\mathbf{M}.\mathbb{D}.\mathbb{D}^T] \\ \mathbb{TR}[\mathbf{M}.\mathbb{G}.\mathbb{G}^T] \end{pmatrix} = \begin{pmatrix} \mathbb{T}_{PP} & \mathbb{T}_{DP} & \mathbb{T}_{GP} \\ \mathbb{T}_{PD} & \mathbb{T}_{DD} & \mathbb{T}_{GD} \\ \mathbb{T}_{PG} & \mathbb{T}_{DG} & \mathbb{T}_{GG} \end{pmatrix} . \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \tag{6.96}$$

$$\text{i.e. } \underline{\alpha} = \underline{\mathbb{T}}_M = \mathbb{T}_{PDG}.\underline{\lambda} \tag{6.97}$$

where $\mathbb{T}_{PDG}$ is the 3×3 matrix in (6.95) whose elements are the matrix traces $\mathbb{T}_{XY}$ defined, using (6.95), by:

$$\mathbb{T}_{XY} = \mathbb{TR}[\mathbf{M}_X.\mathbb{Y}.\mathbb{Y}^T] \tag{6.98}$$

Using (6.97) we now have an expression for the Lagrangian vector, $\underline{\lambda}$, i.e.

$$\underline{\lambda} = \mathbb{T}_{PDG}^{-1}.\underline{\alpha} \tag{6.99}$$

Substituting the expression for $\underline{\lambda}$ given by (6.99) into (6.94) we get, since $\mathbb{T}_{PDG}$ is symmetric,

$$\mathbf{M} = \underline{\alpha}^{T}.\mathbb{T}_{PDG}^{-1}.\mathbf{M}_{PDG} \tag{6.100}$$

Substituting (6.100) into (6.90), we have,

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{TR}[\mathbf{M}.\Sigma_{Gy}] = \alpha_a.\sigma_a^2 + \alpha_u.\sigma_u^2 + \alpha_e.\sigma_e^2 = \sigma^2 \tag{6.101}$$

which shows that the estimator, $\hat{\sigma}^2$, is unbiased what for <u>any</u> matrices, $\mathbf{M}_P$, $\mathbf{M}_D$, and $\mathbf{M}_G$ in (6.94), i.e. <u>unconditionally</u> unbiased.

Using the expression for $\hat{\sigma}^2$ in (6.77), and substituting the value for $\mathbf{M}$ in (6.99), we obtain,

$$\hat{\sigma}^2 = \lambda_1.\mathbb{Q}_P + \lambda_2.\mathbb{Q}_D + \lambda_3.\mathbb{Q}_G = \underline{\alpha}^{T}.\mathbb{T}_{PDG}^{-1}.\underline{\mathbb{Q}}_{PDG} \tag{6.102}$$

where the three element vector $\underline{\mathbb{Q}}_{PDG} = (\mathbb{Q}_P, \mathbb{Q}_D, \mathbb{Q}_G)$ is formed from the quadratic functions whose general form is given, using (6.95), by,

$$\mathbb{Q}_X = \mathbf{y}_T^{T}.\mathbb{G}^{T}.\mathbf{M}_X.\mathbb{G}.\mathbf{y}_T \tag{6.103}$$

From (6.89) we see that (6.102) will produce unconditionally unbiased estimates of σ_a^2, σ_u^2 and σ_e^2 when $\underline{\alpha}^{T}$ is chosen as (1, 0, 0), (0, 1, 0) and (0, 0, 1) respectively. Hence, by substituting these values into (6.101), we find that the three element vector of estimates, $\underline{\hat{\sigma}}^2 = (\hat{\sigma}_a^2, \hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$ is given by:

$$\underline{\hat{\sigma}}^2 = \mathbb{T}_{PDG}^{-1}.\underline{\mathbb{Q}}_{PDG} \tag{6.104}$$

For (6.104) to hold when $(\sigma_a^2, \sigma_u^2, \sigma_e^2) = (\hat{\sigma}_a^2, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$, requires that,

$$\hat{\mathbb{T}}_{PDG}.\underline{\hat{\sigma}}^2 = \underline{\hat{\mathbb{Q}}}_{PDG} \tag{6.105}$$

where $\hat{\mathbb{T}}_{PDG}$ and $= \underline{\hat{\mathbb{Q}}}_{PDG}$ now contain the estimates $\hat{\sigma}_a^2$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ in place of σ_a^2, σ_u^2, σ_e^2.

Incorporating the definition of $\mathbb{T}_{PDG}$, and hence the estimate $\hat{\mathbb{T}}_{PDG}$, in equations (6.96) to (6.98), (6.105) simplifies to:

$$\mathbb{TR}[\hat{\Sigma}_{Gy}^{-1}.\mathbb{P}.\mathbb{P}^T] = \mathbf{y}_T^T.\mathbb{G}^T.\hat{\Sigma}_{Gy}^{-1}.\mathbb{P}.\mathbb{P}^T.\hat{\Sigma}_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \tag{6.106}$$

$$\mathbb{TR}[\hat{\Sigma}_{Gy}^{-1}.\mathbb{D}.\mathbb{D}^T] = \mathbf{y}_T^T.\mathbb{G}^T.\hat{\Sigma}_{Gy}^{-1}.\mathbb{D}.\mathbb{D}^T.\hat{\Sigma}_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \tag{6.107}$$

$$\mathbb{TR}[\hat{\Sigma}_{Gy}^{-1}.\mathbb{G}.\mathbb{G}^T] = \mathbf{y}_T^T.\mathbb{G}^T.\hat{\Sigma}_{Gy}^{-1}.\mathbb{G}.\mathbb{G}^T.\hat{\Sigma}_{Gy}^{-1}.\mathbb{G}.\mathbf{y}_T \tag{6.108}$$

where $\hat{\Sigma}_{Gy}^{-1}$ again contains the residual variance estimates.

Note that equations (6.106) to (6.108) are exactly the same as equations (6.74) to (6.76), i.e. those produced by maximising the likelihood.

From appendix B, (B15) shows that the covariance of two estimators, $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$, is given by a similar result to (6.78), namely,

$$Cov[\hat{\sigma}_i^2, \hat{\sigma}_j^2] = 2.\mathbb{TR}[\mathbf{M}_i.\Sigma_{Gy}.\mathbf{M}_j.\Sigma_{Gy}] \tag{6.109}$$

where $\mathbf{M}_i$ and $\mathbf{M}_j$ are their corresponding matrices as defined in (6.77).

Hence by substituting (6.100) into (6.109), we can show that,

$$Cov[\hat{\sigma}_i^2, \hat{\sigma}_j^2] = 2.\underline{\alpha}_i^T.\mathbb{T}_{PDG}^{-1}.\underline{\alpha}_j \tag{6.110}$$

where $\underline{\alpha}_i$ and $\underline{\alpha}_j$ are the $\underline{\alpha}$ vectors corresponding to $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$.

By substituting different values of $\underline{\alpha}_i$ and $\underline{\alpha}_j$ into (6.110), we obtain the covariance matrix of the vector $\underline{\hat{\sigma}}^2$, in (6.104), as,

$$\mathbb{C}ov[\underline{\hat{\sigma}}^2] = 2.\mathbb{T}^{-1}_{PDG} \tag{6.111}$$

Note from (6.111) that the estimators of the residual variances are not independent.

A simplified version of equation (6.108) may be obtained by multiplying (6.106) by $\hat{\sigma}^2_a$, (6.107) by $\hat{\sigma}^2_u$ and (6.108) by $\hat{\sigma}^2_e$, and then adding. Doing this we get,

$$\mathbb{TR}[\mathbf{I}_{T-d-p}] = T-d-p = \mathbf{y}^T_T.\mathbb{G}^T.\hat{\Sigma}^{-1}_{Gy}.\mathbb{G}.\mathbf{y}_T \tag{6.112}$$

Also, by using equations (6.67) and (6.85), we can write equations (6.106), (6.107) and (6.112), respectively as:

$$\hat{\sigma}^2_e.\mathbb{TR}[\mathbb{P}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{P}] = \mathbf{y}^T_T.\mathbf{D}^T.(\mathbb{P}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{P})^2.\mathbf{D}.\mathbf{y}_T \tag{6.113}$$

$$\hat{\sigma}^2_e.\mathbb{TR}[\mathbb{D}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{D}] = \mathbf{y}^T_T.\mathbf{P}^T.(\mathbb{D}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{D})^2.\mathbf{P}.\mathbf{y}_T \tag{6.114}$$

$$\hat{\sigma}^2_e.(T-d-p) = \mathbf{y}^T_T.\mathbf{D}^T.\mathbb{P}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{P}.\mathbf{D}.\mathbf{y}_T = \mathbf{y}^T_T.\mathbf{P}^T.\mathbb{D}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{D}.\mathbf{P}.\mathbf{y}_T$$

$$= \mathbf{y}^T_T.\mathbb{G}^T.\hat{\Omega}^{-1}_{Gy}.\mathbb{G}.\mathbf{y}_T \tag{6.115}$$

where $\hat{\Omega}_{Gy}$ contains estimated variance ratios.

Notice that the unconditional result of (6.115) above is exactly the same equation as for the conditional result of (6.86) which is reassuringly consistent. It means that should we just happen to choose the optimal "estimated" values for the variance ratios, ω and υ, when producing conditional estimates, we will get the same estimates for the residual variances as if we had used the unconditional approach.

Substituting the value for $\hat{\sigma}_e^2$ in (6.115) into (6.113) and (6.114), we obtain the two equations,

$$\mathbb{TR}[\Omega_P]/\mathbb{TR}[\Omega_P^0] = \mathbf{y}_T^T.\mathbf{D}^T.\Omega_P^2.\mathbf{D}.\mathbf{y}_T/\mathbf{y}_T^T.\mathbf{D}^T.\Omega_P.\mathbf{D}.\mathbf{y}_T \qquad (6.116)$$

$$\text{where } \Omega_P = \mathbb{P}^T.\hat{\Omega}_{Gy}^{-1}.\mathbb{P}$$

$$\mathbb{TR}[\Omega_D]/\mathbb{TR}[\Omega_D^0] = \mathbf{y}_T^T.\mathbf{P}^T.\Omega_D^2.\mathbf{P}.\mathbf{y}_T/\mathbf{y}_T^T.\mathbf{P}^T.\Omega_D.\mathbf{P}.\mathbf{y}_T \qquad (6.117)$$

$$\text{where } \Omega_D = \mathbb{D}^T.\hat{\Omega}_{Gy}^{-1}.\mathbb{D}$$

Equations (6.116) and (6.117) have exactly the same form as equation (5.05) in chapter five, and, although it is not pursued at this point, this would suggest that a similar approach to chapter five could be applied in order to produce an efficient algorithm for their solution.

6.5 CHAPTER SUMMARY

In this chapter we reproduced all the main results of chapters one to four, i.e. sections A and B of stage three in the table in the introduction, on page 10, for the case of seasonality. Rather than list them all, we refer the reader to the contents page for chapter six.

The main point is that the procedures which were applied to estimate the trend values and residual variances required little modification for the seasonal case and again produced analogous results to the those of the non-seasonal case.

Identical trend estimates to those of Generalised Least Squares were produced using Whittaker's Minimisation, (with suitably defined weightings) and State Space methodology, (assuming vague priors).

Also identical residual variance estimates were produced using both Maximum Likelihood and Minimum Variance approaches.

THE ESTIMATION OF AUTOREGRESSIVE PARAMETERS

So far we have assumed that the "d" non-seasonal autoregressive parameters, ϑ_1, ϑ_2, ... , ϑ_d, and the "p" seasonal autoregressive parameters, ϕ_1, ϕ_2, ... , ϕ_p, in the smoothness equations (6.04) and (6.10) of the last chapter are fixed, in the sense that they are pre-specified by the modeller.

We now turn to the case where these parameters are left unspecified and hence need to be estimated. We will refer to this, for convenience, as the "variable" parameter model to distinguish it from the "fixed" parameter models already dealt with, although it should be remembered that we do not mean "variable" in the sense of a random variable, but only in the sense that their true values will be different for different time series.

7.1 LEAST SQUARES ESTIMATION

7.11 THE NON-SEASONAL CASE

In chapter one we observed that from the general form of Whittaker's equation, written for convenience as,

$$\psi = (\mathbf{y}-\mathbf{x})^T.(\mathbf{y}-\mathbf{x}) + (1/\omega).\mathbf{x}^T.\mathbf{D}^T.\mathbf{D}.\mathbf{x} \qquad (7.01)$$

where

$$\mathbf{x} = (x_T, x_{T-1},\ldots,x_1)^T$$
$$\mathbf{y} = (y_T, y_{T-1},\ldots,y_1)^T$$

and **D** was defined by the (T-d) × T matrix,

$$\mathbf{D} = \begin{bmatrix} 1 , -\vartheta_1, -\vartheta_2 \ldots\ldots, -\vartheta_d, 0 , 0 , 0 , \ldots\ldots, 0 \\ 0 , 1 , -\vartheta_1, -\vartheta_2 \ldots\ldots, -\vartheta_d, 0 , 0 , 0 , .., 0 \\ 0 , 0 , \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{etc} \end{bmatrix} \qquad (7.02)$$

we could obtain the least squares estimate of $\mathbf{x}$, i.e. $\hat{\mathbf{x}}$, by differentiating ψ in (7.01) with respect to $\mathbf{x}$ and setting to zero, i.e.

$$\partial\psi/\partial\mathbf{x} = -2.\left[(\mathbf{y}-\hat{\mathbf{x}}) - (1/\omega).\mathbf{D}^T.\mathbf{D}.\hat{\mathbf{x}}\right] = \mathbf{0} \qquad (7.03)$$

which therefore gave $\hat{\mathbf{x}}$, as,

$$\hat{\mathbf{x}} = \left(\mathbf{I}_T + (1/\omega).\mathbf{D}^T.\mathbf{D} \right)^{-1}.\mathbf{y} \qquad (7.04)$$

Similarly, we can also obtain the least squares estimate, $\hat{\vartheta}$, of the parameter vector $\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_d)^T$ by minimising ψ in (7.01) with respect to ϑ, i.e.

$$\partial\psi/\partial\vartheta = \partial(\mathbf{x}^T.\mathbf{D}^T.\mathbf{D}.\mathbf{x})/\partial\vartheta = \partial(\mathbf{a}^T.\mathbf{a})/\partial\vartheta = 0 \qquad (7.05)$$

where $\mathbf{D}.\mathbf{x} = \mathbf{a}$, and $\mathbf{a}$ is the vector of smoothness residuals given by,

$$\mathbf{a} = (a_T, a_{T-1}, \ldots, a_{d+1})^T$$

However, writing out the set of T-d equations, $\mathbf{D}.\mathbf{x} = \mathbf{a}$ in (7.05), in full we have,

$$x_t = \vartheta_1.x_{t-1} + \vartheta_2.x_{t-2} + \ldots + \vartheta_d.x_{t-d} + a_t \qquad \text{for } d < t \leq T \qquad (7.06)$$

which can be alternatively written,

$$\kappa = \mathbb{X}.\vartheta + \mathbf{a} \qquad (7.07)$$

where κ is a (T-d)×1 vector given by,

$$\kappa = (x_T, x_{T-1}, \ldots, x_{d+1})^T \qquad (7.08)$$

and $\mathbb{X}$ is a (T-d)×d matrix given by,

$$\mathbb{X} = \begin{pmatrix} x_{T-1}, & x_{T-2}, & \cdots, & x_{T-d} \\ x_{T-2}, & x_{T-3}, & \cdots, & x_{T-d-1} \\ \cdot & \cdot & \cdots, & \cdot \\ x_{d}, & x_{d-1}, & \cdots, & x_{1} \end{pmatrix} \tag{7.09}$$

The set of equations in (7.06) and (7.07) are in "regression" format and hence their sum of squared errors, $\mathbf{a}^T.\mathbf{a}$, can be minimised giving the usual regression solution,

$$\hat{\vartheta} = (\mathbb{X}^T.\mathbb{X})^{-1}.\mathbb{X}^T.\kappa \tag{7.10}$$

7.111 Algorithms

The joint solutions of (7.04) and (7.10) can be effected by realising that the value of ψ in (7.01) can never be increased by either minimising ψ w.r.t $\mathbf{x}$ for <u>any</u> ϑ using (7.04) or minimising ψ w.r.t. ϑ for <u>any</u> $\mathbf{x}$ using (7.10). Hence we obtain the following algorithm to jointly solve equations (7.04) and (7.10).

1. Choose an initial value for the vector ϑ, say ϑ_0.

2. Form $\mathbf{D}_0$, from ϑ_0, using (7.02).

3. Find $\mathbf{x}_0$ using (7.04), i.e. $\hat{\mathbf{x}}_0 = \left(\mathbf{I}_T + (1/\omega).\mathbf{D}_0^T.\mathbf{D}_0 \right)^{-1}.\mathbf{y}$

4. Form κ_0 and $\mathbb{X}_0$ from $\mathbf{x}_0$ using (7.08) and (7.09).

5. Find a new value for ϑ_0 using (7.10). i.e. $\hat{\vartheta}_0 = (\mathbb{X}_0^T.\mathbb{X}_0)^{-1}.\mathbb{X}_0^T.\kappa_0$.

6. Repeat steps 2 to 5 until convergence where $\hat{\mathbf{x}} = \mathbf{x}_0$, $\hat{\vartheta} = \vartheta_0$.

By repeatedly the above algorithm followed by that described at the end of chapter five, which produces maximum likelihood estimates of

the residual variances, σ_a^2 and σ_e^2, (and hence ω), we can therefore obtain joint estimates of ϑ, $\mathbf{x}$, σ_a^2, σ_e^2, and ω.

7.112 Results

The algorithms described in the last section have been applied to the time series of figure 1.1, (section **1.14**), in chapter one, both for autoregressive lags of d=1 and d=2. The results are shown, respectively, in figures 7.1 and 7.2 which follow.

Trend Estimates
Figure 7.1

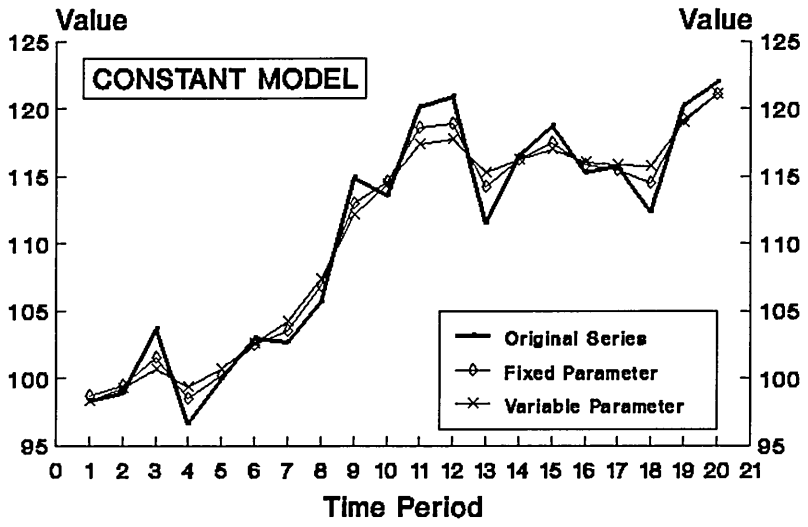


Figure 7.1 shows the results for, what is referred to as the "constant", or lag d=1, model, (in line with the fixed parameter model described in section **5.212** of chapter five, so-called because it was invariant to constant data). As well as the original, ($\mathbf{y}$), series, the estimated trends, ($\hat{\mathbf{x}}$), are plotted for both fixed and variable parameter cases.

As seen from the trend plots in figure 7.1 the variable parameter model gives a slightly smoother trend than that with fixed parameters, although neither are what I would consider to be sufficiently smooth.

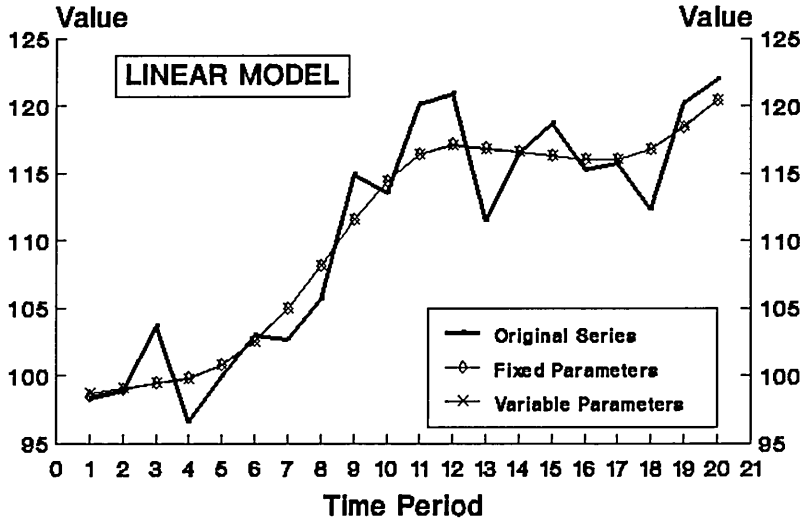
Other estimated values for figure 7.1 are summarised in table 7.1 below.

Table 7.1

Estimates:	ω	σ_e^2	σ_a^2	ϑ_1
Fixed Parameter:	2.4329	5.1813	12.6055	+1.0000
Variable Parameter:	0.9478	7.3705	6.9855	+1.0107

From table 7.1 we see that the residual variance estimates, and in particular the estimate of their ratio, ω, are markedly different. However, when we compare either of these estimates of ω to those used in the examples of figure 1.1, (section **1.14**), of chapter one, we see that neither is small enough to produce enough smoothing, which would seem to require an estimated value of ω of about 0.1, which in turn would suggest that the "constant invariant" or lag d=1 model was inadequate.

Trend Estimates
Figure 7.2



Since the smoothness is, to a certain extent, measured by its residual variance, σ_a^2, and hence ω since $\omega = \sigma_a^2/\sigma_e^2$, smoothness will be

increased by increasing the number of smoothness parameters, ϑ, thereby ensuring a better "fit" for the trend.

Figure 7.2 shows the results of increasing the number of smoothness parameters from one to two, i.e. for, what is referred to as the "linear", or lag d=2, model, (in line with the fixed parameter model described in section **5.212** of chapter five, so-called because it was invariant to linear data). Again, as well as the original, (**y**), series, the estimated trends, ($\hat{\mathbf{x}}$), are plotted for both fixed and variable parameter cases.

Other estimated values for figure 7.2 are summarised in table 7.2 below.

Table 7.2

Estimates:	ω	σ_e^2	σ_a^2	ϑ_1	ϑ_2
Fixed Parameters:	**0.1907**	**9.9298**	**1.8935**	**+2.0000**	**-1.0000**
Variable Parameters:	**0.1797**	**9.9115**	**1.7815**	**+1.8780**	**-0.8761**

In figure 7.2 we see that both the fixed and variable parameter cases produce almost identical trend estimates, $\hat{\mathbf{x}}$; more importantly the estimates now produce, what could fairly be described as smooth series.

This is confirmed by the estimates of ω in table 7.2 which are both of the order of 0.1 coupled with vary similar estimates for the residual variances σ_a^2 and σ_e^2.

It would therefore appear that, on the basis of these results, that a reasonable rule of thumb for model adequacy would be a variance ratio, ω, estimate of about 0.2, although it is recognised that this hardly constitutes even the beginnings of a proper analysis of the topic.

7.12 THE SEASONAL CASE

The least squares estimation of autoregressive parameters for the case of a seasonal model is almost identical to that of the non-seasonal case.

The equivalent seasonal version of Whittaker's equation of (7.01) is given by equation (6.13) of chapter six i.e.

$$\psi = (\mathbf{y}-\mathbf{x}-\mathbf{s})^T.(\mathbf{y}-\mathbf{x}-\mathbf{s}) + (1/\omega).\mathbf{x}^T.\mathbf{D}^T.\mathbf{D}.\mathbf{x} + (1/\upsilon).\mathbf{s}^T.\mathbf{P}^T.\mathbf{P}.\mathbf{s} \qquad (7.11)$$

the only new addition to (7.01) being the vector of seasonal values, $\mathbf{s}$, where $\mathbf{s} = (s_T, s_{T-1}, \ldots . s_1)^T$.

Least squares estimators of $\mathbf{x}$ and $\mathbf{s}$ are given by minimising ψ in (7.11) w.r.t. $\mathbf{x}$ and $\mathbf{s}$ giving equations (6.17) and (6.18) of chapter six, i.e.

$$\hat{\mathbf{x}} = (\Upsilon_T.\Pi_T - \mathbf{I}_T)^{-1}.(\Upsilon_T-\mathbf{I}_T).\mathbf{y} \text{ and } \hat{\mathbf{s}} = (\Pi_T.\Upsilon_T - \mathbf{I}_T)^{-1}.(\Pi_T-\mathbf{I}_T).\mathbf{y}$$

$$\text{where } \Pi_T = \mathbf{I}_T + (1/\omega).\mathbf{D}^T.\mathbf{D} \text{ and } \Upsilon_T = \mathbf{I}_T + (1/\upsilon).\mathbf{P}^T.\mathbf{P} \qquad (7.12)$$

In a similar way to the non-seasonal case, least squares estimators of the autoregressive parameter vector, $\vartheta = (\vartheta_1, \vartheta_2, \ldots . \vartheta_d)^T$ and the seasonally autoregressive parameter vector $\phi = (\phi_1, \phi_2, \ldots . \phi_p)^T$ are given by minimising ψ in (7.11) w.r.t. ϑ and ϕ to give,

$$\hat{\vartheta} = (\mathbb{X}^T.\mathbb{X})^{-1}.\mathbb{X}^T.\kappa \text{ and } \hat{\phi} = (\mathbb{S}^T.\mathbb{S})^{-1}.\mathbb{S}^T.\$ \qquad (7.13)$$

where the formula for $\hat{\vartheta}$, described earlier in the chapter, results from minimising $\mathbf{x}^T.\mathbf{D}^T.\mathbf{D}.\mathbf{x}$ by utilising the regression approach of equations (7.06) and (7.07), thereby defining κ and $\mathbb{X}$ by (7.08) and (7.09); the formula for $\hat{\phi}$ is obtained by exactly the same procedure, $\$$ and $\mathbb{S}$ being the seasonal equivalents to κ and $\mathbb{X}$ about which more will be said in a moment.

Therefore, in exactly the same way as for the non-seasonal case, (7.12) and (7.13) can be evaluated alternatively until convergence to give the least squares estimates $\hat{\mathbf{x}}$, $\hat{\mathbf{s}}$, $\hat{\vartheta}$ and $\hat{\phi}$ which minimise ψ in (7.11).

By utilising the approach of section **6.4** of chapter six to obtain maximum likelihood estimates of the residual variances, in conjunction with the least squares estimation procedure described above, all model parameters can therefore be jointly estimated.

There is a "common sense" proviso to the above procedures. In equations (7.12), the production of the estimates, $\hat{\mathbf{x}}$ and $\hat{\mathbf{s}}$, requires the inversion of the matrix $(\Upsilon_T.\Pi_T-\mathbf{I}_T)$ and its transpose $(\Pi_T.\Upsilon_T-\mathbf{I}_T)$. If, however, the matrices **D** and **P** in (7.11) are identical, it is easily seen from (7.12) that these matrices are singular. Hence to effect a solution **D** and **P** must have a different structure such that the matrix $(\Upsilon_T.\Pi_T-\mathbf{I}_T)$ can be inverted.

What this means in practice is that the seasonal model cannot have the same form as the general trend model of (7.06). In other words we could not employ a general seasonal model of the form,

$$s_t = \phi_1.s_{t-1} + \phi_2.s_{t-2} + \ldots + \phi_p.s_{t-d} + u_t \qquad (7.14)$$

This of course begs the question as to why we would ever want to model seasonality using (7.14) anyway, since there is nothing particularly "seasonal" about it. Hence all the proviso really says is that seasonal models should have some inherently sensible seasonal form.

In this respect, two perfectly acceptable examples of seasonal models, previously defined in equations (6.08) and (6.09) of chapter six, are,

$$s_t = \phi_1.s_{t-s} + \phi_2.s_{t-2s} + \ldots + \phi_p.s_{t-ps} + u_t \qquad (7.15)$$

where $ps < t \leq T$, "s" is the seasonal period and "p" is the seasonal lag, and alternatively,

$$u_t = s_t + s_{t-1} + \ldots + s_{t-s+1} \qquad (7.16)$$

where $s-1 < t \leq T$ and "s" is again the seasonal period.

The relationships (7.15) model the current seasonalities in terms of the seasonalities of previous years. An estimate of ϕ can be obtained from the appropriate regression and it (7.15) can easily be written in the form $\mathbf{P.s = u}$ where $\mathbf{P}$ is an appropriate matrix and $\mathbf{u}$ is a vector of residuals.

Equations (7.16) reflect the idea that sums of yearly seasonalities should be as small as possible and can again be represented by a suitable choice of $\mathbf{P}$ in $\mathbf{P.s = u}$; the estimate of ϕ not being required since the parameters are fixed.

7.2 MAXIMUM LIKELIHOOD ESTIMATION

In this section we look at the estimation of the "d" non-seasonal autoregressive parameters, $\vartheta_1, \vartheta_2, \ldots, \vartheta_d$, and the "p" seasonal autoregressive parameters, $\phi_1, \phi_2, \ldots, \phi_p$, by maximising the likelihood function. In doing so, it should be remembered that we need to introduce the Normality assumption for all residuals since the likelihood function is formed on the basis of that assumption.

7.21 THE NON-SEASONAL CASE

For the non-seasonal case the relevant log-likelihood function is given by equations (4.13) and (4.14) of chapter four i.e.

$$\mathbb{LL}(\omega,\sigma_e^2,\vartheta) = -\Big((T-d).\ln(2\pi) + (T-d).\ln(\sigma_e^2) + \ln|\Omega_\omega| + \mathbf{y}^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y}/\sigma_e^2 \Big)/2 \qquad (7.17)$$

where,

$$\Omega_\omega = \omega.\mathbf{I}_{T-d} + \mathbf{D}.\mathbf{D}^T \qquad (7.18)$$

which, since the vector of parameters, ϑ, is wholly contained in $\mathbf{D}$ and hence Ω_ω, is equivalent to minimising the function, $\mathbb{LL}^*(\omega,\sigma_e^2,\vartheta)$, where,

$$\mathbb{LL}^*(\omega,\sigma_e^2,\vartheta) = \sigma_e^2.\ln|\Omega_\omega| + \mathbf{y}^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y} \qquad (7.19)$$

Whilst it is a straightforward matter to obtain first, (and higher order), derivatives of $\mathbb{LL}^*(\omega,\sigma_e^2,\vartheta)$ with respect to the elements of ϑ, I have not, so far, found it possible to solve the resulting equations when set to zero.

The alternative is to use numerical search techniques, (Scales, 1985; Hamilton, 1994, section 5.7), to obtain the minimum of $\mathbb{LL}^*(\omega,\sigma_e^2,\vartheta)$ in (7.19) and hence the maximum likelihood estimates of ϑ.

For example a multivariate variation on Newton's approximation gives the following iteration sequence,

$$\vartheta^* = \vartheta_0 - [\partial^2\{\mathbb{LL}^*(\omega,\ \sigma_e^2,\ \vartheta_0)\}/\partial\vartheta^2]^{-1}.\partial\{\mathbb{LL}^*(\omega,\ \sigma_e^2,\ \vartheta_0)\}/\partial\vartheta \qquad (7.20)$$

The problem with using (7.20) is that it relies on choosing an initial starting point ϑ_0 which is sufficiently close to the optimum, i.e. within, what might be termed, the, (negative), likelihood "basin"; otherwise (7.20) will not converge.

Substitution of the least squares value of $\mathbf{x}$ in (7.04) into (7.01) gives,

$$\psi(\mathbf{x}=\hat{\mathbf{x}}) = \mathbf{y}^T.\mathbf{D}^T.(\omega.\mathbf{I}_{T-d} + \mathbf{D}.\mathbf{D}^T)^{-1}.\mathbf{D}.\mathbf{y} = \mathbf{y}^T.\mathbf{D}^T.\Omega_\omega^{-1}.\mathbf{D}.\mathbf{y} \qquad (7.21)$$

In other words (7.19) can be written,

$$\mathbb{LL}^*(\omega,\sigma_e^2,\vartheta) = \sigma_e^2.\ln|\Omega_\omega| + \psi(\mathbf{x}=\hat{\mathbf{x}}) \qquad (7.22)$$

Minimisation of the last term, $\psi(\mathbf{x}=\hat{\mathbf{x}})$, of (7.22) was discussed earlier in this chapter and gives the least squares estimates of ϑ. Hence as long as the term $\sigma_e^2.\ln|\Omega_\omega|$ in (7.22) is relatively small, the least squares estimates of ϑ may be used as a starting point for the iteration of (7.20).

In practice this seemed to work very well for "long" series i.e. for series where T >> d. In these cases the likelihood function, (at least for the cases d=1 and d=2, which are the only cases where one can plot the variation of the likelihood against its parameter values), was reasonably well behaved, (in the d=2 case the contours were unimodal and shaped like a stretched ellipse at 45^0 to the parameter axes), with the least squares optimum quite close to that of the likelihood function. Indeed when the length of the time series was much greater than the number of parameters needing to be estimated, i.e. T >>> d, the two optima were almost identical.

The iteration broke down when T and hence T-d was small however, (e.g. T ~ 5 and d ~ 2). In these cases the likelihood surface showed multiple local optima and it appeared to be almost a case of luck as to which optima (7.20) converged to, if indeed it converged at all.

What was worse was that as the value of the variance ratio was varied, the global optimum would "flip" from one of the local optima to the other when particular values of ω were reached.

One may well ask whether the case when T-d is small is important in practice since to model a series of T=5 values say, using d=2 parameters would seem to be gross over-parameterisation. However cases such as T=10 and d=1 are borderline and no analysis would be complete without a full investigation of these limitations and a proper solution to the problem.

Box and Jenkins, (1976, section 7.1, p213), encountered the same problem in their ARIMA modelling in which they also obtain a likelihood function comprised of a determinant term and a sum of squares term similar to (7.22). They say,

" *Usually the* [determinant term] *is of importance only for small* [T]. *For moderate and large values of* [T, the likelihood function] *is dominated by* [the sum of squares term] *and thus the contours of the unconditional sum of squares function in the space of the parameters* (ϕ,ϑ) *are very nearly contours of likelihood and of log-likelihood. It follows, in particular, that the parameter estimates obtained by minimising the sum of squares, which we call least squares estimates, will usually provide very close approximations to the maximum likelihood estimates.....In the remainder of this section and in* [the next section] *our main emphasis will be on the calculation, study, and use of the unconditional sums of squares function...and on calculating least squares estimates.*"

They do not, however, offer any idea as to what *moderate and large values of* T might be, nor do they give any proof that the likelihood function will tend to the sum of squares function as T is increased, which are two issues which need to be addressed.

7.22 THE SEASONAL CASE

Maximum likelihood parameter estimation for the seasonal case follows exactly the same lines as for the non-seasonal case with no extra difficulties.

The log-likelihood function for the seasonal case is given by equations (6.69) and (6.71) of chapter six namely,

$$\mathbb{LL}[\sigma_a^2,\sigma_u^2,\sigma_e^2,\vartheta,\phi] = -1/2.((T-d-p).\ln(2\pi) + \ln|\Sigma_{Gy}| + \mathbf{y}^T.\mathbb{G}^T.\Sigma_{Gy}^{-1}.\mathbb{G}.\mathbf{y}) \quad \ldots.(7.23)$$

where,

$$\Sigma_{Gy} = \mathbb{C}ov[\mathbb{G}.\mathbf{y}_T] = \sigma_a^2.\mathbb{P}.\mathbb{P}^T + \sigma_u^2.\mathbb{D}.\mathbb{D}^T + \sigma_e^2.\mathbb{G}.\mathbb{G}^T \qquad (7.24)$$

Again maximisation of $\mathbb{LL}[\sigma_a^2,\sigma_u^2,\sigma_e^2,\vartheta,\phi]$ in (7.23) is equivalent to minimisation of $\mathbb{LL}^*[\omega,\upsilon,\sigma_e^2,\vartheta,\phi]$ where $\omega = \sigma_a^2/\sigma_e^2$, $\upsilon = \sigma_u^2/\sigma_e^2$ and,

$$\mathbb{LL}^{*}[\omega,\upsilon,\sigma_e^2,\vartheta,\phi] = \sigma_e^2.\ln|\Omega_{Gy}| + \mathbf{y}^T.\mathbb{G}^T.\Omega_{Gy}^{-1}.\mathbb{G}.\mathbf{y} \tag{7.25}$$

$$\text{with } \Omega_{Gy} = \mathbb{G}.\mathbb{G}^T + \upsilon.\mathbb{D}.\mathbb{D}^T + \omega.\mathbb{P}.\mathbb{P}^T \tag{7.26}$$

Using the relationships, $\mathbb{G} = \mathbb{P}.\mathbf{D} = \mathbb{D}.\mathbf{P}$, from equations (6.67) of chapter six, we can write Ω_{Gy} as,

$$\Omega_{Gy} = \mathbb{G}.[\mathbf{I}_T + \upsilon.\Delta_D + \omega.\Delta_P].\mathbb{G}^T \tag{7.27}$$

$$\text{where } \Delta_D = \mathbf{D}.(\mathbf{D}.\mathbf{D})^{-2}.\mathbf{D}^T \text{ and } \Delta_P = \mathbf{P}.(\mathbf{P}.\mathbf{P})^{-2}.\mathbf{P}^T \tag{7.28}$$

and hence,

$$\Omega_{Gy} = \mathbb{G}.\Delta_D.\Omega.\Delta_P.\mathbb{G}^T \tag{7.29}$$

$$\text{where } \Omega = \mathbf{D}^T.\mathbf{D}.\mathbf{P}^T.\mathbf{P} + \upsilon.\mathbf{D}^T.\mathbf{D} + \omega.\mathbf{P}^T.\mathbf{P} \tag{7.30}$$

Also substitution of (7.12) into (7.11) gives the seasonal equivalent to (7.21), namely,

$$\psi(\mathbf{x}=\hat{\mathbf{x}},\mathbf{s}=\hat{\mathbf{s}}) = \mathbf{y}^T.\mathbf{P}^T.\mathbf{P}.\Omega^{-1}.\mathbf{D}^T.\mathbf{D}.\mathbf{y} \tag{7.31}$$

Hence to show that the sum of squares term, i.e. $\mathbf{y}^T.\mathbb{G}^T.\Omega_{Gy}^{-1}.\mathbb{G}.\mathbf{y}$, in (7.25) is equal to $\psi(\mathbf{x}=\hat{\mathbf{x}},\mathbf{s}=\hat{\mathbf{s}})$ in (7.31), we need, using (7.29), to show that,

$$\mathbb{G}^T.[\mathbb{G}.\Delta_D.\Omega.\Delta_P.\mathbb{G}^T]^{-1}.\mathbb{G} = \mathbf{P}^T.\mathbf{P}.\Omega^{-1}.\mathbf{D}^T.\mathbf{D} \tag{7.32}$$

At this point in time (7.32) is still a conjecture, although the matrix equality has been simulated several times and on each occasion all corresponding matrix elements were identical.

The point of knowing (7.32) to be true is, of course, that we may then use least squares estimates of ϑ and ϕ as a starting point for the minimisation of (7.25) using a seasonally equivalent iteration to (7.20).

7.3 CHAPTER SUMMARY

This chapter has been concerned exclusively with the estimation of the autoregressive parameters for both the non-seasonal and seasonal models, i.e. the "General", variable parameter, models and relates to section C of stage three in the table on page 10.

The two estimation procedures investigated were Least Squares and Maximum Likelihood, which produced different estimates although it looks possible, (although this is still a conjecture), that these may be asymptotically, (i.e. for long time series), the same. The advantages of using Least Squares estimates in preference to their "more optimal" Likelihood counterparts is twofold. Firstly, they can be produced more more efficiently, and secondly, they are not subject to the problems of having to differentiate between local and global optima. However the area needs further investigation to be able to properly clarify the situation.

FURTHER DEVELOPMENTS

During the course of a thesis, and, in particular this thesis, one is constantly aware of the many areas which could be, but are not being, fully investigated and, to a great extent, this results from a simple matter of priorities. This is not to say, however, that one does not spend some time thinking how they might be approached and whether to do so would seem to be a relatively straightforward matter or whether one envisages a whole host of obstacles.

In this chapter we give an indication of what other areas need to be dealt with and also indicate how, on the face of it, these might be addressed.

8.1 LIMITING MODELS

It is interesting to observe what happens when one, (or more), of the residual variances, and hence their corresponding set of residuals, tends to zero.

8.11 THE NON-SEASONAL CASES

The general two non-seasonal model equations were:

$$y_t = x_t + e_t \tag{8.01}$$

$$x_t = \vartheta_1.x_{t-1} + \vartheta_2.x_{t-2} + \ldots + \vartheta_d.x_{t-d} + a_t \tag{8.02}$$

8.111 Letting σ_e^2 Tend To Zero

As σ_e^2, and hence each e_t in (8.01), tends to zero, y_t tends to x_t in

(8.01) and hence (8.02) becomes simply,

$$y_t = \vartheta_1.y_{t-1} + \vartheta_2.y_{t-2} + \ldots + \vartheta_d.y_{t-d} + a_t \qquad (8.02)$$

For a variable parameter model, the autoregressive parameter vector, ϑ, and σ_a^2 can be estimated from (8.02) using ordinary regression on previous values of y_t.

For a fixed parameter, (i.e. ϑ pre-specified), model, (8.02) effectively becomes,

$$y_t = \mu_t + a_t \qquad (8.03)$$

where μ_t is known, and all that remains is the estimation of the mean and variance of the a_t using conventional methods.

8.112 Letting σ_a^2 Tend To Zero.

As σ_a^2, and hence each a_t in (8.02), tends to zero, (8.02) becomes,

$$x_t = \vartheta_1.x_{t-1} + \vartheta_2.x_{t-2} + \ldots + \vartheta_d.x_{t-d} \qquad (8.04)$$

We observe, by inspection, that the solution to (8.04) is given by:

$$x_t = \beta_1.m_1^t + \beta_2.m_2^t + \ldots + \beta_d.m_d^t \qquad (8.05)$$

where each m_i, for i=1 to d, is a root of the polynomial,

$$m_i^d - \vartheta_1.m_i^{d-1} - \vartheta_2.m_i^{d-2} - \ldots - \vartheta_{d-1}.m_i - \vartheta_d = 0 \qquad (8.06)$$

Note that in the event that two of the roots of (8.06) are equal, e.g.

$m_1 = m_2 = m_j$, we would replace the terms $\beta_1.m_1^t + \beta_2.m_2^t$ in (8.05) by $\beta_1.m_j^t + \beta_2.t.m_j^t$.

Similarly, in the event that three of the roots of (8.06) are equal, e.g. $m_3 = m_4 = m_5 = m_k$, we would replace the terms $\beta_3.m_3^t + \beta_4.m_4^t + \beta_5.m_5^t$ in (8.05) by $\beta_3.m_k^t + \beta_4.t.m_k^t + \beta_5.t^2.m_k^t$ and so on for other sets of equal roots.

Also, in the event that two roots of (8.06), m_u and m_v, form a complex pair, say $r.e^{i\alpha}$ and $r.e^{-i\alpha}$, the corresponding two terms in (8.05) will be $\beta_u.r^t.\cos(\alpha.t)$ and $\beta_v.r^t.\sin(\alpha.t)$.

Substituting (8.05) into (8.01) gives,

$$y_t = \beta_1.m_1^t + \beta_2.m_2^t + \ldots + \beta_d.m_d^t + e_t \qquad (8.07)$$

For the fixed parameter model the m_i, i=1 to d, could be found from (8.06) and hence the β_j, j=1 to d, and an estimate of σ_e^2 from (8.07) using ordinary regression.

For the variable parameter model, initial estimates of the ϑ_k, k=1 to d, could be found by substituting $y_t - e_t$, (with e_t=0 to begin with), for x_t in (8.04) and applying ordinary regression. This would enable the m_i, i=1 to d, to be found from (8.06) and hence the β_j, j=1 to d, and estimates of e_t and σ_e^2 from (8.07) using ordinary regression. This would then allow better estimates of ϑ_k to be found and so on.

A quicker, but perhaps not so interesting, way of producing the same estimates is, using the matrix notation of equation (7.01) of the previous chapter, to minimise,

$$(\mathbf{y}-\mathbf{x})^T.(\mathbf{y}-\mathbf{x}) + \lambda^T.\mathbf{D}.\mathbf{x} \qquad (8.08)$$

where λ is a T-d vector of lagrangian multipliers, which has solution,

$$\mathbf{x} = (\mathbf{I}_T + \mathbf{D}^T.(\mathbf{D}.\mathbf{D}^T)^{-1}.\mathbf{D}).\mathbf{y} \qquad (8.09)$$

and hence gives the fixed parameter solution.

To obtain the variable parameter solution, (8.09) can be used in conjunction with the regression of (8.04), which produces estimates for the vector of autoregressive parameters, ϑ, given by equation (7.10) of the previous chapter i.e.

$$\hat{\vartheta} = (\mathbb{X}^T.\mathbb{X})^{-1}.\mathbb{X}^T.\kappa \qquad (8.10)$$

By feeding the results of (8.10) into (8.09) and vice versa until convergence we obtain the required estimates of $\mathbf{x}$ and ϑ. Substitution of (8.09) into (8.08) then gives us an estimate of σ_e^2 i.e.

$$\hat{\sigma}_e^2 = \mathbf{y}^T.\mathbf{D}^T.(\mathbf{D}.\mathbf{D}^T)^{-1}.\mathbf{D}.\mathbf{y}/T \qquad (8.11)$$

8.12 THE SEASONAL CASES

The non-seasonal case is easily extended to include seasonality using the ideas of the above sections. The only reason we do not deal with them here is that it would mean detailing twelve variations in all i.e. the combination of models produced by letting either one or two of the residual variances σ_e^2, σ_a^2 and σ_u^2 tend to zero and considering both their fixed and variable parameter versions.

8.2 EXTENDING THE MODEL

We showed how the non-seasonal model could be extended to deal with seasonality, and, in doing so, saw that this was easily accommodated within the general framework and hence presented no particularly new problems. We now consider two further areas which could be

accommodated.

8.21 BUSINESS CYCLES

As far as the model is concerned the "so-called" business cycle is just another, extra, form of seasonality and can be dealt with in exactly the same way.

8.22 STEP CHANGES

Perhaps my main concern with the model as it stands, is its ability to deal with step changes in a time series or, what are sometimes referred to as, "interventions", (see Box and Tiao, 1975, McCleary and Hay, 1980 and Cleary and Levenbach, 1982). These are changes in the level of a series over a specified length of time resulting from some form of outside intervention, for example, as a result of a sales promotion or a tax change.

However this is easily accommodated within the measurement equation thus, (for the seasonal model),

$$y_t = x_t + s_t + c_t + e_t \qquad (8.12)$$

where for a single change, for example, c_t would have values of "c" for the period of the change and zero otherwise.

The estimation of the parameter "c" is then found by either minimising the appropriate form of Whittaker's function with respect to "c", (i.e. least squares estimation), or by maximising the appropriate likelihood function, in exactly the same way that autoregressive parameters were estimated in chapter seven.

8.3 MODEL ADEQUACY

Two possible criteria for judging a particular model's adequacy are the extent to which (a) the autoregressive parameters and (b) the residual variances are significantly different from zero.

8.31 AUTOREGRESSIVE PARAMETER SIGNIFICANCE

For fixed parameter models this boils down to whether the autoregressive lags are the correct ones to use, and was addressed by Akaike, (see section **0.24** of the introduction). For variable parameter models we need to use the likelihood ratio test for parameter inclusion (see appendix). Both of these tests are asymptotic and hence are only really valid for relatively long time series, which begs the question as to whether we need to go to the bother of calculating maximum likelihood rather than least squares parameter estimates since, as it would appear to be the case, (see chapter seven), that the latter tend to the former for long time series.

8.32 RESIDUAL VARIANCE SIGNIFICANCE

In section **8.1** of this chapter we began to look at what would happen as one or more of the residual variances tended to zero. Again to test this we could use the asymptotic likelihood ratio test, (see appendix), although whether a more satisfactory test could be produced by considering their estimates from the point of view of quadratic forms rather than maximum likelihood, (see chapter four), is worth pursuing further.

For example, I am able to show that, for the fixed parameter model, under the hypothesis $\hat{\omega} = \omega$, the statistic $(T-1).\hat{\sigma}_a^2 / \sigma_a^2$ has a chi-squared distribution with T-1 degrees of freedom, although the search for two such, independent statistics, which would permit a meaningful test, has not proved successful so far.

8.4 FORECASTING

For fixed parameter models, the area of forecasting has already been extensively dealt by utilising the State Space approach of chapter two. It is, after all, essentially what the "prediction" stage, (see section **2.4** of chapter two), is concerned with and will produce forecasts and their "errors" quite happily.

For the variable parameter case, things are not so straightforward and we must tread very carefully, since, for example, the forecasting of x_{t+1} and hence y_{t+1} using equations (8.01) and (8.02), leads to all sorts of problems with regard to the correlations between the trend, **x**, and parameter, ϑ, estimates.and ϑ_1 etc.

8.5 DATA TRIALS

This thesis has mostly concerned itself with developing the theory behind a suggested "trend" model. As such there has not been time to give the model the rigorous testing it needs by applying it to wide selection of different time series with different characteristics, using both real and simulated data.

In the case of simulated data this is done to test to what extent the model can replicate the parameters of the original series, and, has been done to some extent, when used to test the truth of a derived theoretical result.

However, the approach was neither structured nor systematic and certainly did not lay down any performance criteria by which a model could be judged. This was even more true for "real life" data; the model having only been applied to about five series, whose adequacy was only judged by eye. In this respect the collection of "1001" series, (Makridakis, 1982), would be a good starting point for testing both model adequacy and forecasting ability.

8.6 MULTIVARIATE MODELS

In section **0.144** of the introduction we described the property of "Additivity" and how it might be tested as long as the trend could be written in the form of equation (0.12). This was the first and last time that "Additivity" was mentioned and as such I should, perhaps, say a little more about how I would envisage this area could be further developed.

The first point to note is that the trend model developed in the these does indeed fit the requirement of equation (0.12) and as such we could use the ideas of section **8.4** to test the adequacy of this property both for real and simulated series.

However, the idea of several "disaggregated" series being consistent with their "aggregated" counterpart can be taken a stage further by asking whether, instead of treating both the "disaggregated" and the "aggregated" series as essentially separate, and then applying the results of the univariate trend model so far described to each series in the hope of obtaining a match, we cannot build the additive relationship, (using lagrangian multipliers say), into a joint multivariate model, which would ensure such a property.

It must be said that the above is about as far as my thinking has gone in this area. However, given the essentially linear constraint of additivity, I would imagine there was every hope of success. In fact, why stop at "Additivity"; since if it worked, it would only seem a small step to building a general multivariate linear relationship into such models whose parameters could be either fixed or variable.

CONCLUSION

In considering the achievements and limitations of this thesis, we should perhaps begin by reminding ourselves of the objectives which we set down in the introduction.

Our original intention was to develop an approach which would estimate the trend based on an initial set of "desirable" and, more to the point, unambiguously "definable", (preferably in mathematical terms), properties. As a starting point, four properties and their initial definitions were offered, namely "Fidelity", "Smoothness", "Invariance" and "Additivity", which I felt trend estimates should inherently possess. Hence , in this critique, we should, perhaps, begin by looking at the extent to which, on the one hand, the approach incorporated these properties and, on the other, the resulting trend displayed them.

Accepting the limited number of example data sets which were considered, the trend's "Fidelity" showed no obvious weaknesses, since its definition easily accommodated the problems of seasonality and, at least on the face of it, appeared to be flexible enough to be able to be extended further to the areas of business cycles and interventions without any foreseeable difficulties.

Similarly the criterion of "Smoothness" slotted nicely into the scheme of things, and was again easily extended to include seasonality. Moreover we were encouraged to see that the resulting trend estimates, at least for correctly specified models, did indeed follow a smooth path.

The only negative feature, (of incorporating the "Smoothness" definition), arose in the estimation of the autoregressive parameters for the "General", (variable parameter), model, which, although, straightforward for the case of Least Squares estimates, looked less than satisfactory for those produced using Maximum Likelihood.

Nevertheless, even here, there were indications that both types of estimate would be identical for relatively long time series, which is exactly the criterion that would need to be met before any of the advantageous inferential properties of likelihood estimates could be utilised.

A luxury of the model was that the "Invariance" property would already be accommodated within the definitions of "Fidelity" and "Smoothness", at no extra cost so to speak, its only drawback being that the general model would appear to break down if truly invariant data, (i.e. conforming exactly to a simple polynomial), were used. However the breakdown would only be to one of the model's simpler limiting forms, which had consequently simpler estimation procedures.

The "Additivity" property of the trend estimates has not been tested so far. However, because of their linear nature, they did possess a form which would allow this property to be, not only tested, but hopefully extended, to a multivariate model in which "Additivity" could be in-built as a special case.

Having considered the properties of the estimates, we also need to address, what might be termed, the "effectiveness" of the estimation procedures, themselves, and by this we mean both effective in the sense that:

(a) they can be reasonably efficiently carried out, (since there is little point in producing a model which, for all practical purposes, is "inestimable"), and

(b) they in some sense produce the "best" estimates.

In the case of (a), (with the exception of the Likelihood smoothness parameter estimates discussed earlier), the calculation of estimates has been developed, (and in some cases explored in some depth), and efficient procedures subsequently produced, which have then gone on to be programmed to give example results, whose accuracy and speed of execution compares favourably with, say, those of other statistical

techniques such as regression.

In considering the extent to which these are "best" estimates, as in (b), we have adopted a slightly more pragmatic approach to the normal one of considering the estimators' properties and distributions. The problem here is that because of the non-linear relationships between the unknowns, especially, in the general model, it is very unlikely that any estimators will have any of the more commonly used distributions, and, in common with most non-linear models, will, at best, only have asymptotically limiting forms. Hence, rather than following this track, we have deliberately limited our search to estimators that are produced as a result of some form of "optimal" process, (on the assumption that this will result in estimators which will possess "optimal" properties).

In doing so, however, we have tried to look at as many optimal processes as possible, and have not been happy to accept an estimator, unless we can show that it results from at least two such optimal processes, which, at least initially, seem to be based on different criteria. Thus in the case of the trend estimates, we have shown that we get the same estimates from Whittaker's function minimisation, Generalised Least Squares regression and State Space estimation. The same residual variance estimates were produced whether we used Maximum Likelihood or Minimum Variance, (of quadratic forms), estimation, and, it would appear that, (for long series, anyway), the Least Squares and Maximum Likelihood smoothness parameter estimates will be identical.

Moreover, there are other desirable bi-products arising from having investigated more than one estimation procedure for each estimate.

Firstly it leads us to ask whether the two procedures will produce the same estimates for other models, and, in general, to find the set of models for which this is the case and hence why this should be the case? For example, can any model capable of being expressed as a General Regression model also be formulated in State Space format, and will their respective estimates always be identical under the assumption of vague prior information.

Secondly, one procedure often, by its nature, reveals estimator properties which the other does not and vice-versa. For example, the estimation of residual variances using Maximum Likelihood does not, (except asymptotically), give us the estimators' variances, which the minimum variance approach produces. Also the dual estimation approach of the autoregressive parameters suggests that Least Squares estimates will have the same asymptotic Normal distributions as do Maximum Likelihood estimates.

Lastly, the "more the merrier" approach to estimation also produces a pay-off when it comes to practical implementation. See, for example, the way the Minimum Variance and Maximum Likelihood approaches were able to complement each other in producing an efficient algorithm to find their common estimates.

Whether these estimates have other desirable properties or convenient distributions from which efficient "inference" tests can be produced is, we would argue, not within the terms of reference of this thesis and has therefore not been pursued, as neither have, for that matter, the problems of forecasting, (after all forecastability was not one of our initial "desirable" trend properties). In this sense the purpose of this thesis was to develop something akin to the Least Squares Estimation component of the Regression model, leaving, as Legendre did in 1805, its fuller implications and consequences to others.

In conclusion therefore, this thesis offers an alternative approach to trend estimation which is direct, clearly defined and easily executed.

Abraham B. and J. Ledolter (1983). *Statistical Methods for Forecasting.* Wiley.

Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle.* 2nd Int. Symp. on Information Theory B. N. Petrov and F. Csaki eds. Akademiai Kiado, Budapest.

Akaike H. (1974). *A new look at the statistical model identification.* IEEE Trans. on Auto. Control AC-19 pp 716-723.

Akaike H. (1979a). *A Bayesian extension of the minimum AIC procedure of autoregressive model fitting.* Biometrika 66: 237-242.

Akaike H. (1979b). *On the construction of composite time series models.* Proc. 42nd Session Int. Stat. Inst.

Akaike H. (1979c). *Smoothness priors and the distributed lag estimator.* Technical Report 40 (T. A. Anderson, Project Director), Department of Statistics, Stanford University, Stanford, Calif.

Akaike H. (1980a). *Likelihood and the Bayes procedure.* Bayesian statistics. Proceedings of the 1st International Meeting. ed. Bernado J. M. et al, University Press, Valencia, Spain, p 143-203.

Akaike H. and M. Ishiguro (1980). *BAYSEA: A Bayesian seasonal adjustment program.* Tokyo: The Institute of Statistical Mathematics, Computer Science Monographs No. 13.

Akaike H. (1980b). *Seasonal adjustment by a Bayesian modelling.* J. Time Series Anal. 1(1):1-13.

Akaike H. (1980c). *Likelihood of a model and information criteria.* Paper presented at the Conference on Model Selection April 18-21, Gainesville, Florida.

Anderson B. D. O. and J. B. Moore (1979). *Optimal filtering.* Englewood Cliffs: Prentice-Hall.

Ansley C. F. and R. Kohn (1982). *A geometrical derivation of the fixed interval smoothing algorithm.* Biometrika 69: 486-7.

Bowerman B. L. and R. T. O'Connell (1979). *Forecasting and Time Series.* PWS (Wadsworth) Publishers.

Box G. and G. Jenkins (1976). *Time Series Analysis, Forecasting and Control.* Holden-Day, Oakland, Calif.

Box G. and G. Tiao (1975). *Intervention Analysis with applications to Economic and Environmental Problems.* Journal of American Statist. Assoc, vol 70, p70-.

Brotherton T. and W. Gersch (1981). *A Data Analytic approach to the Smoothing Problem and some of its variations.* Proceedings of the 20th IEEE Conference on Decision and Control, pp 1061-1069.

Brown R. G. (1959). *Statistical Forecasting for Inventory Control.* McGraw-Hill, New York.

Brown R. G. (1962). *Smoothing, Forecasting and Prediction of Discrete Time Series Data.* Prentice-Hall, New Jersey.

Bureau of the Census (1969). *X-11 Information for the user.* U. S. Department of Commerce, Washington, DC, U. S. Government Printing Office.

Cleary J. P. and H. Levenbach (1982). *The Professional Forecaster.* Belmont, Calif. Lietime Learning Publications.

Cramer J. S. (1989). *Econometric Applications of Maximum Likelihood Methods.* Cambridge University Press.

Craven P. and G. Wahba (1979). *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalised cross-validation.* Numer. Math. 31: 377-403.

Dagum E. B. (1975). *Seasonal Factor forecasts from ARIMA models.* Proceedings of the 40th session of the International Statistical Institute, Warsaw, Poland vol 3 pp 206-219.

Dagum E. B. (1980). *The X-11 ARIMA seasonal adjustment method.* Statistics, Canada, Seasonal Adjustment and Time Series Staff, Research Paper, Ottawa.

Davidson R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics.* Oxford University Press, New York.

Den Butter F. A. G. and M. M. G. Fase (1991). *Seasonal Adjustment as a practical problem.* North-Holland.

Dhrymes P. J. (1970). *Econometrics. Statistical foundations and applications.* Harper and Row.

Enders W. (1995). *Applied Econometric Time Series.* Wiley.

Eubank R. L. (1986). *A note on smoothness priors and non-linear regression.* J. Amer. Stat. Assoc. 81(394): 514-517.

Gersch W. and G. Kitagawa (1983a). *A time varying multivariate autoregressive modelling of econometric time series.* Stat. Res. Divison, Bureau of the Census, U. S. Dept. Commerce, Tech Paper 9.

Gersch W. and G. Kitagawa (1983b). *The prediction of time series with trends and seasonalities.* J. Bus. and Econ. Stats. 1(3): 253-264.

Gersch W. and G. Kitagawa (1984). *Transfer function estimation: A smoothness priors approach.* Proceedings of the 23rd IEEE Conference on Decision and Control pp 363-368.

Gersch W. (1987). *Some applications of smoothness priors in time series.* Proceedings of 26th IEEE Conference on Decision and Control pp 1684-1689.

Gersch W. and G. Kitagawa (1988). *Smoothness priors in time series.* Bayesian Analysis of Time Series and Dynamic Models, ed. J. C. Spall, Marcel Dekker Inc. pp 431-476.

Gersch W. and G. Kitagawa (1989). *Smoothness priors transfer-function estimation.* Automatica 25(4): 603-608.

Gersch W. (1989). *Smoothness priors multi-channel autoregressive time series modelling.* International Conference on Acoustics, Speech and Signal Processing vol 4: 2158-2161 (IEEE cat. no. 89CH2673-2).

Gersch W. and D. Stone (1990). *Multi-channel time varying autoregressive modelling: a circular lattice-smoothness priors realisation.* Proceedings of 29th IEEE Conference on Decision and Control vol 2: 859-60 (cat. no. 90CH2917-3).

Golub G., M. Heath and G. Wahba (1979). *Generalised cross-validation as a method for choosing a good ridge parameter.* Techometrics 21: 215-223.

Good I. J. (1965). *The estimation of probabilities.* M. I. T. Press, Cambridge, Mass. p 75-76.

Good I. J. (1963). *Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables.* Annals of Mathematical Statistics. 34: 911-934.

Good I. J. (1971a). *The probabilistic explication of information, evidence, surprise, causality, explanation, utility, (with appendix, discussion and replies).* Foundations of Statistical Inference. Toronto: Holt, Reinhart and Winston of Canada. pp. 108-141.

Good I. J. (1971b). *Non-parametric roughness penalty for probability densities.* Nature Physical Science. 229: 29-30.

Good I. J. and R. A. Gaskins (1971). *Non-parametric roughness penalties for probability densities.* Biometrika. 58: 255-277.

Good I. J. and R. A. Gaskins (1972). *Global non-parametric estimation of probability densities.* Virginia Journal of Science. 23: 171-193.

Good I. J. and R. A. Gaskins (1980). *Density estimation and bump hunting by the penalised likelihood method exemplified by scattering of meteorite data.* J. Amer. Stat. Ass. 75:42-73.

Graybill F. A. (1969). *Matrices with applications in statistics.* Wadsworth Int.

Hamilton J. D. (1994). *Time Series Analysis.* Princeton University Press.

Harrison P. J. and C. F. Stevens (1976). *Bayesian Forecasting.* J. Roy. Statist. Soc., Series B, Vol. 38, pp 205-228.

Harvey A. C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge University Press.

Healy M. J. R. (1991). *Matrices for Statistics.* Clarendon Press, Oxford.

Hylleberg S. (1986). *Seasonality in Regression.* New York, Academic Press.

Jazwinski A. H. (1970). *Stochastic processes and filtering theory.* New York: Academic Press.

Kalman R. E. (1960). *A new approach to linear filtering and prediction problems.* J. of Basic Engineering, Transactions ASME. Series D 82: 35-45.

Kalman R. E. (1963). *New Methods in Wiener Filtering Theory.* Proceedings of the First Symposium of Engineering Applications of Random Function Theory and Probability eds., Wiley pp. 270-388.

Kendall M. (1976). *Time-Series.* Griffin, p 29.

Kimmeldorf G. S. and G. Wahba (1970). *A correspondance between Bayesian estimation on stochastic processes and smoothing by splines.* Annals of Mathematical Statistics 41: 495-502.

Kimmeldorf G. S. and G. Wahba (1971). *Some results on Tchebycheffian spline functions.* Journal of Mathematical Analysis and Applications 38: 82-105.

Kitagawa G. and H. Akaike (1978). *A procedure for the modelling of non-stationary time series.* Ann. Inst. Stat. Math. 30-B:351-363.

Kitagawa G. (1981). *A non-stationary time series model and its fitting by a recursive filter.* J. Time Series Anal. 2(2):103-116.

Kitagawa G. and W. Gersch (1982). *A smoothness priors approach to the modelling of time series with trends and seasonalities.* Amer. Statist. Assoc. Proc. of the Bus. and Econ. section pp 403-408.

Kitagawa G. and W. Gersch (1984). *A smoothness priors-state space modelling of time series with trend and seasonality.* J. Amer. Statist. Assoc. 79(386): 378-389.

Kitagawa G. and W. Gersch (1985a). *A smoothness priors time-varying AR coefficient modelling of non-stationary covariance time series.* IEEE Transactions on Automatic Control AC-30(1): 48-56.

Kitagawa G. and W. Gersch (1985b). *A smoothness priors long AR model method for spectral estimation.* IEEE Transactions on Automatic Control AC-30(1): 57-65.

Kohn R. and C. F. Ansley (1988). *Equivalence between Bayesian smoothness priors and optimal smoothing for function estimation.* Bayesian Analysis of Time Series and Dynamic Models, ed. J. C. Spall, Marcel Dekker Inc. pp 393-430.

Lindley D. V. (1965b). *Introduction to Probability and Statistics from a Bayesian point of view. Part one. Probability.* Cambridge University Press.

Lindley D. V. (1965a). *Introduction to Probability and Statistics from a Bayesian point of view. Part two. Inference.* Cambridge University Press.

Lindley D. V. (1971). *Making Decisions.* Wiley.

Lindley D. V. and A. F. M. Smith (1972) *Bayes estimates for the linear model.* J. R. Stat. Soc. Ser. B(34): 1-41.

Magnus J. R. and H. Neudecker (1995). *Matrix Differential Calculus with applications in Statistics and Econometrics.* Wiley.

McCleary R. and R. A. Hay (1980). *Jr. Applied Time Series Anal. for the Social Sciences.* Sage Publications.

Makridakis S. et al (1982). *The accuracy of extrapolation (time series) methods: results of a forecasting competition.* J. Forecasting 1:111-153.

Meditch J. S. (1969). *Stochastic optimal linear estimation and control.* McGraw Hill.

O'Muircheartaigh C. and D. P. Francis (1981). *Statistics, a dictionary of terms and ideas.* Arrow Books.

Polasek W. (1990). *Vector distributed lag models with smoothness priors.* Computational Statistics and Data Analysis 10(2): 133-141.

Priestly M. B. (1981). *Spectral Analysis and Time Series.* Academic Press.

Reinsch C. H. (1967). *Smoothing by spline functions.* Numer. Math. 10: 177-183.

Sage A.P. and J.L. Melsa (1971). *Estimation theory with applications to communications and control.* New York: McGraw-Hill.

Scales L. E. (1985). *Introduction to Non-Linear Optimisation.* Macmillan.

Schlicht E. (1981). *A seasonal adjustment principle and a seasonal adjustment method derived from this principle.* J. Amer. Stat. Assoc. 76: 374-378.

Searle S. R. (1971). *Linear Models.* J. Wiley & sons.

Searle S. R. (1982). *Matrix Algebra useful for statistics.* J. Wiley & sons.

Shiller R. J. (1973). *A distributed lag estimator derived from smoothness priors.* Econometrica 41(4): 775-788.

Shiller R. J. (1984). *Smoothness priors and non-linear regression.* J. Amer. Stat. Assoc. 79(387): 609-615.

Shishkin J., A. H. Young and J. C. Musgrave (1967). *The X-11 variant of the Census Method-II Seasonal Adjustment Program.* Tech. Paper No. 15, U. S. Department of Commerce, Bureau of the Census, Washigton, DC.

Taylor W. E. (1974). *Smoothness priors and stochastic prior restrictions in distributed lag estimation.* International Economic Review. 15: 803-804.

Terasvirta T., G. Yi and G. Judge (1988). *Model selection, smoothing and parameter estimation in linear models under squared error loss.* Computational Statistics Quarterly 4(3): 191-205.

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

Theil H. and A. S. Goldberger (1961). *On pure and mixed statistical estimation in economics.* International Economic review, 2: 65-78.

Thurman S. S., P. A. O. B. Swamy and J. S. Mehta (1986). *An examination of distributed lag model coefficients estimated with smoothness priors.* Communications in Statistics-Theory and Methods 15(6): 1723-1749.

Wahba G. and S. Wold (1975a). *A completely automatic french curve: fitting spline functions by cross-validation.* Communications in Statistics. 4: 1-17.

Wahba G. and S. Wold (1975b). *Periodic splines for spectral density estimation: the use of cross-validation for determining the correct degree of smoothing.* Communications in Statistics. 4: 125-141.

Wahba G. (1977). *A survey of some smoothing problems and the method of cross-validation for solving them.* Applications of Statistics, Amsterdam, North Holland, ed. P. R. Krishnaiah, 507-524.

Wahba G. (1978). *Improper priors, spline smoothing and the problem of guarding against model errors in regression.* Journal of the Royal Statistical Society. Series B. 40: 364-372.

Weinert H. L. (1979). *Statistical methods in optimal curve fitting.* Communications in Statistics. 87: 525-536.

Whittaker E. T. (1923). *On a new method of graduation.* Proc. Edinburgh Math. Soc., 63-75.

Whittaker E. T. and G. Robinson (1924). *The Calculus of Observations*, Blackie and Son Ltd. Chapter XI: "Graduation, or the smoothing of data", 285-315.

Winters P. R. (1960). *Forecasting Sales by Exponentially Weighted Moving Averages.* Management Science, Vol. 6, No. 3, pp 324-342.

Woonacott T. H. and R. J. Woonacott (1976). *Introductory Statistics.* Wiley.

APPENDIX A

LINEAR, UNBIASED ESTIMATORS

FOR FIXED PARAMETERS

The usual form of the Gauss-Markov theorem refers to the equation system,

$$\underline{y} = \mathbf{X}.\underline{\beta} + \underline{e} \qquad (A1)$$

where $\underline{y}$ is a T×1 vector of observations, $\mathbf{X}$ is a T×m matrix of known constants, $\underline{e}$ is a T×1 vector of residuals coming from a distribution whose mean is zero and whose covariance matrix $\sigma_e^2.\Xi$ is also known, and $\underline{\beta}$ is a m×1 vector of unknown parameters, i.e. $\underline{\beta}^T = \{\beta_1, \beta_2, \ldots\ldots \beta_m\}$.

Under these conditions we seek a linear estimator $\hat{\beta}_i$ of parameter β_i, given by:

$$\hat{\beta}_i = \mathbf{m}_i^T.\underline{Y}, \quad \text{where } \mathbf{m}_i^T \text{ is a } 1\times T \text{ vector} \qquad (A2)$$

The criterion for choosing the vector $\mathbf{m}_i$ is that the resulting variance of $\hat{\beta}_i$ is to be a minimum.

Hence we need to minimise,

$$\mathbb{V}\left[\hat{\beta}_i\right] = \mathbb{E}\left[\left(\hat{\beta}_i - \mathbb{E}[\hat{\beta}_i]\right)^2\right] = \mathbb{E}\left[\left(\mathbf{m}_i^T(\underline{Y} - \mathbb{E}[\underline{Y}])\right)^2\right]$$

$$= \mathbb{E}\left[\left(\mathbf{m}_i^T.(\underline{Y} - \mathbf{X}.\underline{\beta})\right)^2\right] = \mathbb{E}\left[\left(\mathbf{m}_i^T.\underline{E}\right)^2\right]$$

$$= \mathbf{m}_i^T.\mathbb{E}\left[\underline{E}.\underline{E}^T\right].\mathbf{m}_i = \sigma_e^2.\mathbf{m}_i^T.\Xi.\mathbf{m}_i \qquad (A3)$$

using (A1), (A2), and the fact that $\sigma_e^2.\Xi$ was defined as the covariance matrix of $\underline{E}$.

We impose the condition that the estimator is unbiased, i.e.

$$\mathbb{E}[\hat{\beta}_i] = \beta_i$$

or since $\mathbb{E}[\hat{\beta}_i] = \mathbb{E}[\mathbf{m}_i^T.\underline{Y}] = \mathbf{m}_i^T.\mathbb{E}[\underline{Y}] = \mathbf{m}_i^T.\mathbb{E}[\mathbf{X}.\underline{\beta} + \underline{E}] = \mathbf{m}_i^T.\mathbf{X}.\underline{\beta}$

$$\text{that} \quad \mathbf{m}_i^T.\mathbf{X}.\underline{\beta} = \beta_i = \mathbf{v}_i^T.\underline{\beta}, \quad \text{implying that} \quad \mathbf{m}_i^T.\mathbf{X} = \mathbf{v}_i^T \qquad \text{(A4)}$$

where $\mathbf{v}_i$ is an m×1 vector whose elements are all zero except for the 'i' th which equals unity.

To include condition (A4) in the minimisation we need to minimise the function $f(\mathbf{m}_i)$, where:

$$f(\mathbf{m}_i) = \sigma_e^2.\mathbf{m}_i^T.\Xi.\mathbf{m}_i - (\mathbf{m}_i^T.\mathbf{X} - \mathbf{v}_i^T).\underline{\lambda} \qquad \text{(A5)}$$

with respect to the vector $\mathbf{m}_i$, where $\underline{\lambda}$ is an m×1 vector of Lagrangian multipliers.

Hence,

$$\partial f(\mathbf{m}_i)/\partial \mathbf{m}_i = 2.\sigma_e^2.\Xi.\mathbf{m}_i - \mathbf{X}.\underline{\lambda}$$

i.e.

$$\mathbf{m}_i = 1/2\sigma_e^2.\Xi^{-1}.\mathbf{X}.\underline{\lambda} \qquad \text{(A6)}$$

Using (A6) with condition (A4), we have,

$$\mathbf{v}_i = \mathbf{X}^T.\mathbf{m}_i = 1/2\sigma_e^2.\mathbf{X}^T.\Xi^{-1}.\mathbf{X}.\underline{\lambda}$$

which gives,

$$\underline{\lambda} = 2.\sigma_e^2.\left(X^T.\Xi^{-1}.X \right)^{-1}.v_i$$

which we can now substitute into (A6) to give,

$$m_i = \Xi^{-1}.X.\left(X^T.\Xi^{-1}.X \right)^{-1}.v_i \qquad (A7)$$

Further substitution of (A7) into (A2) obtains the estimator

$$\hat{\beta}_i = v_i^T.\left(X^T.\Xi^{-1}.X \right)^{-1}.X^T.\Xi^{-1}.\underline{Y} \qquad (A8)$$

By writing equations (A8) for each i in vector form, we obtain the unbiased linear minimum variance estimator, $\underline{\hat{\beta}}$, sometimes referred to as the best linear unbiased estimator, (BLUE), of $\underline{\beta}$ as:

$$\underline{\hat{\beta}} = \left[X^T.\Xi^{-1}.X \right]^{-1}.X^T.\Xi^{-1}.\underline{Y} \qquad (A9)$$

Hence its covariance matrix of is given by:

$$\mathbb{C}ov[\underline{\hat{\beta}}] = \left[X^T.\Xi^{-1}.X \right]^{-1}.X^T.\Xi^{-1}.\mathbb{C}ov[\underline{Y}].\Xi^{-1}.X.\left[X^T.\Xi^{-1}.X \right]^{-1}$$

which since $\mathbb{C}ov[\underline{Y}] = \sigma_e^2.\Xi$ leads to:

$$\mathbb{C}ov[\underline{\hat{\beta}}] = \sigma_e^2.\left[X^T.\Xi^{-1}.X \right]^{-1} \qquad (A10)$$

It can also be noted that for Normally distributed $\underline{E}$, the estimator is also a Maximum Likelihood (MLE) estimator, since it is easily shown that it maximises the likelihood of $\underline{Y}$ by minimising the exponent of the probability density of $\underline{Y}$ with respect to $\underline{\beta}$, i.e.

$$\left(\underline{y} - \mathbf{X}.\underline{\beta} \right)^{T}.\Xi^{-1}.\left(\underline{y} - \mathbf{X}.\underline{\beta} \right) \;=\; \underline{e}^{T}.\Xi^{-1}.\underline{e} \qquad \text{(A11)}$$

Because it minimises the sum of squares in (A11), it can also be regarded as the Generalised Least Squares, (GLS), estimator of $\underline{\beta}$, (minimising the sum of squared "standardised" errors. A proof is provided, by splitting (A3) into the sum of two squares, in (*Wonnacott and Wonnacott,1979*), although it is more easily achieved by simply differentiating (A11) by $\underline{\beta}$.

FOR STOCHASTIC PARAMETERS

The extension is to the case where $\underline{\beta}$, in (A1), is stochastic in the sense that it is randomly drawn from a prior distribution whose mean is $\bar{\underline{\beta}}$. The surprising result is that not only do the features and formula of (A9) still hold for an estimator, $\hat{\underline{\beta}}$ of $\bar{\underline{\beta}}$, but that $\hat{\underline{\beta}}$ can also be regarded as the closest estimator to the vector variate $\underline{\beta}$ itself, in that it minimises the mean squared error of all combinations of $\hat{\underline{\beta}}$ and $\bar{\underline{\beta}}$.

The proof follows almost exactly the same lines as for the fixed parameter case. Equation (A1) still defines the estimator but condition (A3) is achieved by the condition that the estimator is what is known as unconditionally unbiased, (*u-unbiased*), i.e. its vector mean equals $\bar{\underline{\beta}}$, the vector mean of $\underline{\beta}$, i.e.

$$\mathbb{E}\left[\hat{\beta}_i \right] \;=\; \mathbb{E}\left[\beta_i \right] \;=\; \bar{\beta}_i, \qquad \text{for any i} \qquad \text{(A12)}$$

Hence,

$$\mathbb{E}[\hat{\beta}_i] = \mathbb{E}[\mathbf{m}_i^{T}.\underline{Y}] = \mathbf{m}_i^{T}.\mathbb{E}[\underline{Y}] = \mathbf{m}_i^{T}.\mathbb{E}[\mathbf{X}.\underline{\beta} + \underline{E}] = \mathbf{m}_i^{T}.\mathbf{X}.\mathbb{E}[\underline{\beta}] = \mathbf{m}_i^{T}.\mathbf{X}.\bar{\underline{\beta}}$$

which using (A12) gives:

$$\mathbb{E}[\beta_i] = \bar{\beta}_i = \mathbf{m}_i^{T}.\mathbf{X}.\bar{\underline{\beta}} = \mathbf{v}_i^{T}.\bar{\underline{\beta}}, \qquad \text{implying that} \qquad \mathbf{m}_i^{T}.\mathbf{X} = \mathbf{v}_i^{T}$$

At this stage we need to minimise the mean squared error of $\hat{\beta}_i$, then the equivalent first line to (A3) becomes:

$$MSE\left[\hat{\beta}_i \right] = \mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2 \right]$$

$$= \mathbb{E}\left[\left(\underline{m}_i^T.(\underline{Y} - \mathbf{X}.\underline{\beta})\right)^2\right] = \mathbb{E}\left[\left(\underline{m}_i^T.\underline{E}\right)^2\right] \qquad (A13)$$

The proof is identical from that point and the resulting estimator is given by equation (A9) and referred to as the minimum mean square linear u-unbiased estimator (MMSULE), i.e.

$$\underline{\hat{\beta}} = \left[\mathbf{X}^T.\Xi^{-1}.\mathbf{X} \right]^{-1}.\mathbf{X}^T.\Xi^{-1}.\underline{Y} \qquad (A14)$$

Similarly, its covariance matrix of is given by:

$$\mathbb{C}ov[\underline{\hat{\beta}}] = \left[\mathbf{X}^T.\Xi^{-1}.\mathbf{X} \right]^{-1}.\mathbf{X}^T.\Xi^{-1}.\mathbb{C}ov[\underline{Y}].\Xi^{-1}.\mathbf{X}.\left[\mathbf{X}^T.\Xi^{-1}.\mathbf{X} \right]^{-1}$$

which since $\mathbb{C}ov[\underline{Y}] = \mathbf{X}.\mathbb{C}ov[\underline{\beta}].\mathbf{X}^T + \sigma_e^2.\Xi$ leads to:

$$\mathbb{C}ov[\underline{\hat{\beta}}] = \sigma_e^2.\left[\mathbf{X}^T.\Xi^{-1}.\mathbf{X} \right]^{-1} + \mathbb{C}ov[\underline{\beta}] \qquad (A15)$$

Also continuing (A13) on the lines of (A4) it is easily shown that the matrix,

$$\sigma_e^2.\left[\mathbf{X}^T.\Xi^{-1}.\mathbf{X} \right]^{-1} \qquad (A16)$$

has diagonal elements,

$$\mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2 \right] = MSE\left[\hat{\beta}_i \right]$$

and off-diagonal elements,

$$\mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right).\left(\hat{\beta}_j - \beta_j\right) \right]$$

In other words the matrix in (A16) can be regarded as a matrix of mean squared errors and can be written:

$$MSE[\underline{\hat{\beta}}] = \sigma_e^2.\left[\mathbf{X}^T.\Xi^{-1}.\mathbf{X} \right]^{-1} \qquad (A17)$$

with (A15) now being written as:

$$\mathbb{C}ov[\underline{\hat{\beta}}] = MSE[\underline{\hat{\beta}}] + \mathbb{C}ov[\underline{\beta}] \qquad (A18)$$

Note that just as for the fixed parameter case the two results are distribution independent and that if the relevant distribution of E_t is Normal, we can drop the restriction on linearity and apply the results of (A11).

APPENDIX B

THE MEAN, VARIANCE AND COVARIANCE OF QUADRATIC FORMS

THE MEAN AND VARIANCE

Standard results from multivariate theory, demonstrated by (*Searle,1971,chapter 2*), state that:

If the random vector **Z** has a Multivariate Normal distribution given by,

$$\mathbf{Z} \sim \mathbf{N}\left(\varnothing;\ \Sigma \right) \tag{B1}$$

where $\varnothing$ is its zero mean vector and Σ is its square covariance matrix, then, the quadratic form $\mathbf{Z}^T.\mathbf{A}.\mathbf{Z}$, where **A** is a square symmetrical matrix of constants, has a:

(i) Mean given by the trace of the matrix product $\Sigma.\mathbf{A}$, i.e.

$$\mathbb{E}[\mathbf{Z}^T.\mathbf{A}.\mathbf{Z}] = \mathbb{TR}[\Sigma.\mathbf{A}] \tag{B2}$$

(In fact this result is distribution free)

(ii) Variance given by the trace of the matrix product $\Sigma.\mathbf{A}.\Sigma.\mathbf{A}$, i.e.

$$\mathbb{V}[\mathbf{Z}^T.\mathbf{A}.\mathbf{Z}] = 2.\mathbb{TR}[\Sigma.\mathbf{A}.\Sigma.\mathbf{A}] \tag{B3}$$

We now extend these two results to the case to the quadratic form $\mathbf{Z}^T.\mathbf{M}.\mathbf{Z}$, where **M** is square but not necessarily symmetric.

Since $Z^T.M.Z$ is scalar and hence symmetric,

$$Z^T.M.Z = (Z^T.M.Z)^T = Z^T.M^T.Z \qquad (B4)$$

Therefore,

$$Z^T.M.Z = (Z^T.M.Z + Z^T.M^T.Z)/2 = Z^T.(M + M^T).Z/2 \qquad (B5)$$

But $(M + M^T)/2$ is symmetric and hence we can apply equation (B.02) to give,

$$\mathbb{E}[Z^T.M.Z] = \mathbb{E}[Z^T.(M+M^T).Z]/2 = \mathbb{TR}[\Sigma.(M+M^T)]/2 \qquad (B6)$$

And from the properties of the trace of a matrix,

$$\mathbb{TR}[\Sigma.M] = \mathbb{TR}[M.\Sigma] = \mathbb{TR}[\Sigma.M^T] = \mathbb{TR}[M^T.\Sigma] \qquad (B7)$$

Hence,

$$\mathbb{E}[Z^T.M.Z] = \mathbb{TR}[\Sigma.M] \qquad (B8)$$

proving that equation (B2) is also true for non-symmetric matrices.

Similarly, applying equation (B3), we get,

$$\mathbb{V}[Z^T.M.Z] = \mathbb{V}[Z^T.(M+M^T).Z]/4 = 2.\mathbb{TR}[\Sigma.(M+M^T).\Sigma.(M+M^T)]/4 \qquad (B9)$$

which simplifies after expanding and using (B7) to,

$$\mathbb{V}[Z^T.M.Z] = 2.\mathbb{TR}[\Sigma.M.\Sigma.M] \qquad (B10)$$

proving that equation (B3) is also true for non-symmetric matrices.

THE COVARIANCE

We can now extend the result of (B10) to find the covariance of two estimators, $Z^T.M_1.Z$ and $Z^T.M_2.Z$.

$$\mathbb{V}[Z^T.M_1.Z + Z^T.M_2.Z] = \mathbb{V}[Z^T.(M_1 + M_2).Z]$$

$$= \mathbb{V}[Z^T.M_1.Z] + \mathbb{V}[Z^T.M_2.Z] + 2.\mathbb{C}ov[Z^T.M_1.Z,\ Z^T.M_2.Z] \qquad (B11)$$

But, from (B10),

$$\mathbb{V}[Z^T.(M_1 + M_2).Z] = 2.\mathbb{T}\mathbb{R}[\Sigma.(M_1 + M_2).\Sigma.(M_1 + M_2)] \qquad (B12)$$

$$\mathbb{V}[Z^T.M_1.Z] = 2.\mathbb{T}\mathbb{R}[\Sigma.M_1.\Sigma.M_2)] \qquad (B13)$$

$$\mathbb{V}[Z^T.M_2.Z] = 2.\mathbb{T}\mathbb{R}[\Sigma.M_2.\Sigma.M_2] \qquad (B14)$$

Hence, substituting (B12) to (B14) into (B11), $\mathbb{C}ov[Z^T.M_1.Z,\ Z^T.M_2.Z]$ is given by,

$$\mathbb{T}\mathbb{R}[\Sigma.(M_1 + M_2).\Sigma.(M_1 + M_2)] - \mathbb{T}\mathbb{R}[\Sigma.M_1.\Sigma.M_2)] - \mathbb{T}\mathbb{R}[\Sigma.M_2.\Sigma.M_2]$$

$$= \mathbb{T}\mathbb{R}[\Sigma.(M_1 + M_2).\Sigma.(M_1 + M_2) - \Sigma.M_1.\Sigma.M_2 - \Sigma.M_2.\Sigma.M_2]$$

$$= \mathbb{T}\mathbb{R}[\Sigma.M_1.\Sigma.M_2 + \Sigma.M_2.\Sigma.M_1] = 2.\mathbb{T}\mathbb{R}[\Sigma.M_1.\Sigma.M_2]$$

We therefore obtain the result,

$$\mathbb{C}ov[Z^T.M_1.Z,\ \ Z^T.M_2.Z] = 2.\mathbb{T}\mathbb{R}[\Sigma.M_1.\Sigma.M_2] \qquad (B15)$$

APPENDIX C

PARTIAL DIFFERENTIATION OF MATRIX FUNCTIONS BY MATRICES

It often proves necessary to partially differentiate a scalar function of the individual elements of a matrix by each of these elements in turn. If the matrix is a rectangular m×n matrix, the result is m×n expressions, which are then usually set to zero since the reason for such a differentiation is normally to find an optimum.

These m×n expressions are most conveniently expressed as an m×n matrix themselves, with each element corresponding to the result of differentiating the scalar function by the corresponding element of the original matrix, (see Magnus, 1995 for a full treatment).

Suppose that the m×n matrix **X** has ij th element x_{ij} and that y is a function of all these elements i.e. $y = f(x_{ij};\ i = 1,..m;\ j = 1,..n)$, which we can abbreviate to $y = f(\mathbf{X})$. The result of partially differentiating the function y by each element of **X** in turn can be conveniently represented by another m×n matrix, say **Z**, whose elements z_{ij} are given by:

$$z_{ij} = \partial y/\partial x_{ij} \qquad (C1)$$

In matrix terms this can be written as:

$$\partial y/\partial \mathbf{X} = \partial f(\mathbf{X})/\partial \mathbf{X} = \mathbf{Z} \qquad (C2)$$

Particular instances of this technique used in this chapter relate to the case when the scalar function is the trace of a matrix. The three results that are needed are:

1. $\underline{\partial \mathrm{TR}[\mathbf{X}]/\partial \mathbf{X}}$

Here, **X** must obviously be square, i.e. m × m say, and its trace is given by,

$$\mathbb{TR}[\mathbf{X}] = \sum_{i=1}^{m} x_{ii} \qquad (C3)$$

Hence $\partial\mathbb{TR}[\mathbf{X}]/\partial x_{ij}$ equals 1 if i=j and 0 if i≠j, and therefore,

$$\boxed{\partial\mathbb{TR}[\mathbf{X}]/\partial\mathbf{X} = \mathbf{I}_m} \qquad (C4)$$

2. $\underline{\partial\mathbb{TR}[\mathbf{X}.\mathbf{A}]/\partial\mathbf{X}}$

If the matrix **X** is dimensioned m×n then **A**, assumed to be a matrix of constants, must have dimensions n×m.

$$\mathbb{TR}[\mathbf{X}.\mathbf{A}] = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}.a_{ji} \qquad (C5)$$

Hence $\partial\mathbb{TR}[\mathbf{X}.\mathbf{A}]/\partial x_{ij} = a_{ji}$ and therefore

$$\boxed{\partial\mathbb{TR}[\mathbf{X}.\mathbf{A}]/\partial\mathbf{X} = \mathbf{A}^T} \qquad (C6)$$

which simply equals **A** if **A** and therefore **X** is square and **A** is symmetric.

3. $\underline{\partial\mathbb{TR}[\mathbf{X}.\mathbf{A}.\mathbf{X}.\mathbf{A}]/\partial\mathbf{X}}$

Here **X** and **A** must obviously be square and

$$\mathbb{TR}[\mathbf{X.A.X.A}] = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} x_{ij}.a_{jk}.x_{kl}.a_{li} \qquad \text{(C7)}$$

Hence $\partial\mathbb{TR}[\mathbf{X.A.X.A}]/\partial\mathbf{X} = 2.\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{n} a_{jk}.x_{kl}.a_{li}$ and therefore

$$\boxed{\partial\mathbb{TR}[\mathbf{X.A.X.A}]/\partial\mathbf{X} = 2.[\mathbf{A.X.A}]^{T} = 2.\mathbf{A}^{T}.\mathbf{X}^{T}.\mathbf{A}^{T}} \qquad \text{(C8)}$$

which simply equals $2.\mathbf{A}.\mathbf{X}^{T}.\mathbf{A}$ if **A** is symmetric

APPENDIX D

THE LIKELIHOOD RATIO TEST FOR INCLUSION OF PARAMETERS

If L(+r) is the maximum log-likelihood value with the inclusion of r "extra" parameters, and L is the maximum log-likelihood when these parameters are omitted, then under the hypothesis that the additional r parameters are all zero, then the statistic,

$$2(L(+r)-L)$$

has an asymptotic χ^2 distribution with r degrees of freedom.

Note also that any vector, $\hat{\Omega}$, of M.L. estimates of Ω is asymptotically Normally distributed with mean Ω and covariance matrix $-[\partial^2 L/\partial\Omega^2]^{-1}(\Omega=\hat{\Omega})$ which permits standard errors of $\hat{\Omega}$ to be calculated, and hence Z-values for use in identification of significant parameters.