Diagnostics in time series analysis

Warnes, Alexis

Alexis Warnes

Diagnostics in Time Series Analysis

The portmanteau diagnostic test for goodness of model fit is studied. It is found that the true variances of the estimated residual autocorrelation function are potentially deflated considerably below their asymptotic level, and exhibit high correlations with each other. This suggests a new portmanteau test, ignoring the first $p + q$ residual autocorrelation terms and hence approximating the asymptotic chi-squared distribution more closely. Simulations show that this alternative portmanteau test produces greater accuracy in its estimated significance levels, especially in small samples.

Theory and discussions follow, pertaining to both the Dynamic Linear Model and the Bayesian method of forecasting. The concept of long-term equivalence is defined.

The difficulties with the discounting approach in the DLM are then illustrated through an example, before deriving equations for the step-ahead forecast distribution which could, instead, be used to estimate the evolution variance matrix $\mathbf{W}_t$. Non-uniqueness of $\mathbf{W}$ in the constant time series DLM is the principal drawback with this idea; however, it is proven that in any class of long-term equivalent models only p degrees of freedom can be fixed in $\mathbf{W}$, leading to a potentially diagonal form for this matrix.

The bias in the k^{th} step-ahead forecast error produced by any TSDLM variance (mis)specification is calculated. This yields the variances and covariances of the forecast error distribution; given sample estimates of these, it proves possible to solve equations arising from these calculations both for V and p elements of $\mathbf{W}$. Simulations, and a "head-to-head" comparison, for the frequently-applied steady model illustrate the accuracy of the predictive calculations, both in the convergence properties of the sample (co)variances, and the estimates $\widehat{V}$ and $\widehat{W}$. The method is then applied to a 2-dimensional constant TSDLM. Further simulations illustrate the success of the approach in producing accurate on-line estimates for the true variance specifications within this widely-used model.

Diagnostics in Time Series Analysis

1994

Alexis Warnes

Ph.D.

University of Durham

Department of Mathematical Sciences

Contents

I declare that no part of this thesis has been submitted for a degree at any university other than the University of Durham.

Acknowledgements

These are the people without whom I would have *really* struggled...

Peter Craig, my supervisor and troubleshooter, whom I have yet to ask a statistical question to which he hasn't the answer;
Vernon Armitage, and his ever-present moral (and academic) guidance;
Jonty, and his LaTeX skills;
Tom, and his bakery data set;
Catherine and Croc, and cryptic crosswords;
Campbell, and his ramblings.

My mother too, not just for never-ending support, but the great motivation from her unbounded enthusiasm for meeting Sir Peter Ustinov. My father; again, not just for support and enthusiasm, but for providing a much-needed competitive edge through his book-writing successes. And Lucy, for teaching me what a split-infinitive is.

Introduction

"The truth is rarely pure, and never simple"
Oscar Wilde

Nearly all non-mathematicians have a fundamental difficulty visualising the feasibility of research in a mathematical field - namely the preconception that everything must be either right or wrong in a completely black or white manner. Hence the only way it is seen to be possible to research with any form of originality is through advancing the great boundaries of 'truth'. This requires the brilliant application of compellingly innovative ideas, and is a gift which only a handful of intellectual geniuses have ever possessed. Luckily for most of us, there is another vast ocean of originality on which we can set sail, which was so succinctly expressed by Oscar Wilde: the impure waters between truth and opinion, fact and assumption.

It is truly daunting to see the mass of literature written on an area such as Box-Jenkins time series analysis, and realise that, unless you are one of the extremely rare talents who can reshape both the foundational methodology and structural applications of an entire school of thought, you can do little besides merely applying what you have just read, hopefully solving the quest for innovation *en route* by undertaking the analysis of some 'original' data. But the very search for an original data set - moreover, one whose meaningful analysis is

feasible to attempt in a three year project - is almost as difficult as developing the original approach. Consequently, one soon construes that the way forward is through a search of the murkier waters of any statistical method or system, and learns to stop reading abruptly at any mention of 'assumption', 'approximately' or (even better) 'in our opinion', before attempting either to lighten or darken the shade of grey that exists beneath each of these phrases. And, occasionally, one's explorations lead to the discovery of possible improvements, or, more gratifyingly (and far more excitingly), into an alternative approach/solution altogether; this latter course led to the production of Chapter 4 in entirety for this thesis.

Contents

The first opportunity for exploration came when reading about the two alternative versions of the portmanteau test statistic: the original proposed by Box and Pierce in 1970, with an improvement suggested by Ljung and Box in 1978. These diagnostic procedures have a common assumption between them - that the residual autocorrelation function terms, $\hat{r}_k$ for $k = 1, 2, \ldots$, are all independent and Normally distributed, hence allowing us to compare the sum of the squares of m of these sample $\hat{r}_k$s against an appropriate χ^2 distribution.

However, by examining the actual means and variances of the residual autocorrelations (for various AR(p) processes) in section 1.4, as well as the correlations between them at different lags, two conclusions are drawn. Firstly, the true variances of the $\hat{r}_k$ terms are deflated considerably below their assumed

asymptotic values, not only for smaller lags k, which was postulated by Box and Pierce, but also for much higher lags, implying that care is required even up to $k = 10$, and not just $k = p$, if the absolute size of the sample residual acf is to be used as a diagnostic tool. Secondly, it is apparent from derivations in section 1.5 that the first $p + q$ of these $\hat{r}_k$ terms are the most removed from their assumed asymptotic distribution, and exhibit the highest correlations with one another. Given that, when dealing with small sample sizes ($n \simeq 50$), the relevant χ^2 distributions are invariably poor as approximations of the true distributions for both existing test statistics - the original Box-Pierce's S and Ljung-Box's S' - it is shown in section 1.6 that by ignoring these first $p + q$ autocorrelation function terms we produce a third alternative statistic, S'', which fits the appropriate χ^2 distribution far more closely in these small sample sizes. The theoretical calculations of the chapter are then backed up through extensive simulations in section 1.7.

Chapter 2 begins with a brief motivation for the major change in focus that occurred shortly after completing work on the portmanteau statistic. This motivation is best summarised here through recollection of my first encounters with Box-Jenkins methodology. In 1989, whilst working on my undergraduate summer project, there came a moment of great excitement when I took a series of monthly temperatures that had been recorded at my home for several years, and could firstly deseasonalise them, then calculate the estimated and partial autocorrelation functions of the deseasonalised remainder, decide which ARIMA(p, d, q) process to fit, promptly do so, and hence predict the next year's weather over Loughborough (such are the misguided joys for a naive and overea-

ger practitioner...). My father - and I have never forgotten how casually he deflated my enthusiasm - simply made two statements: firstly, that he could have made my predictions by hand (B.F.E. syndrome again), and secondly that he couldn't see what physical interpretation my ARIMA model had with respect to this deseasonalised data. A few years on and I can vaguely justify an AR(1) model fitting such data (warm months generally follow warm months, although this is a gross simplification and misrepresentation of a global warming issue!), but I still find it difficult to argue rudimentarily with the first comment...

There are many frustrations which underlie this story, all of which are highlighted further in section 2.1. These frustrations were to remain buried for a couple of years after completion of my first project, from which point I will forever be indebted to the foresight of my supervisor (and the influences he came under), who towards the end of my first year as a postgraduate succeeded in prising me away from the comforting black-box methodology of Box-Jenkins analysis, and presented me with an untouched version of *Bayesian Forecasting and Dynamic Models* by Mike West and Jeff Harrison. Up until that moment, my experience of Bayesian methods had been limited to the simple applied probability examples of Bayes theorem. Thus, to be confronted with West and Harrison's ideas was akin to learning a new language; one which soon revealed itself to be much richer, allowing the speaker the freedom that I had been starting to feel deprived of under the Box-Jenkins tongue.

The majority of chapter 2 is hence concerned with building the foundations and structure of the Bayesian approach to forecasting, through development of the Dynamic Linear Model (DLM) in section 2.2, and introduction of the

Kalman Filter (section 2.3). Exponential families and conjugate priors are discussed before moving on to a detailed example of the non-linear DLM in section 2.5. This advertising awareness example - together with much of chapter 2 - is taken directly from West and Harrison's book (and related papers), and is useful not only as motivation and clarification, but also since it makes use of the discounting approach in forming a posterior value for the state vector from the given prior (which is the subject of section 2.6); moreover, it is made use of again in chapter 3. A specific class of DLM, the Time Series DLM, is defined in section 2.7, with the concepts of observability and canonical equivalence defined in sections 2.8 and 2.9. Finally, chapter 2 is closed with a definition of long-term equivalence (a concept defined by West and Harrison as general equivalence), referred to extensively later in the thesis.

This change of approach (if not direction, since the goals of time series analysis must surely remain constant, whichever language you speak) in the course of this thesis provided a wealth of not only assumptions, but opinions too (indeed, it is difficult to perform a full Bayesian analysis without expressing a belief of one kind or another...). One of the more salient 'opinions' that must be formed in an analysis is on the appropriateness of the discounting method for loss of information from posterior to prior, and so the importance of this debate is firstly motivated in section 3.1, then the weaknesses of discounting highlighted in section 3.2. A more serious flaw, the paradoxical inability of the discounting approach to work in the presence of full (or very accurate) knowledge of one or more of the elements in the state vector, is then illustrated through a simulated extension to the earlier advertising awareness example. It is also shown how a

more careful choice of an additive form of loss of information - namely through defining the state evolution variance matrix $\mathbf{W}_t$ - results in a far more accurate adaptation of the model to changes in the underlying state.

After the further discussions of section 3.4, the care evidently required in this choice of $\mathbf{W}_t$, and the uncertainty associated with it, motivates a possible method for estimating this variance matrix more accurately, and in an on-line manner, which is derived in section 3.5. However, this method utilises the step-ahead forecast distribution, and although it appears theoretically possible to calculate $\mathbf{W}_t$ fully from the on-line estimates of this distribution, it is further shown in section 3.6 that in the constant TSDLM, $\mathbf{W}$ (non-scalar) is always overparamatrised with respect to the forecast function $f_t(k)$; so much so, in fact, that it is not possible to define $\mathbf{W}$ uniquely beyond its diagonal elements. It is then proven in theorem 3.3, via two lemmata, that the class-defining state evolution variance matrix, $\mathbf{W}$, of two long-term equivalent models (i.e. which have identical forecast functions), is always reducible to a diagonal form (as long as the variance matrix form for $\mathbf{W}$ is still satisfied). Ultimately in this chapter, in section 3.7, the two aforementioned options for modelling the sequential loss of information from posterior to prior are compared and contrasted.

Given that the practitioner is interested in fully or partially specifying the variance matrix $\mathbf{W}$ - which, together with the scalar observational variance V, is all that he need specify once a particular TSDLM is chosen as representing the data evolution adequately - it is not sufficient simply to quote the step-ahead forecast distribution as containing all the information that he needs, and leave him to extract that information for himself. At the beginning of chapter 4, there

is a discussion on the dangers of sub-optimal filtering. This method has been employed by all previous authors who have indeed simply cited the forecast distribution with a shout of 'Eureka!', and then proceeded to advocate feeding back estimates of this distribution directly to the equations arising from the Kalman Filter, in an attempt to solve for both the observational variance V, and the state evolution variance matrix $\mathbf{W}$. This sub-optimal filtering technique has a fundamental flaw - when either V or $\mathbf{W}$ (or, more likely, both) are misspecified, there are potentially large biases introduced into the estimates of the step-ahead variances and covariances within the forecast distribution. By feeding some of these estimates directly back into the Kalman Filter, and taking no account of the biases therein, large errors are transmitted forward immediately into the resultant estimates of the variances, $\widehat{V}$ and $\widehat{\mathbf{W}}$. These biases in the estimated forecast distribution - caused by the very misspecifications we are trying to correct - can hence easily lead to divergences in $\widehat{V}$ and $\widehat{\mathbf{W}}$.

These problems with sub-optimal filtering are so intrinsically related to the methods themselves that they cannot be paid the lip-service they have received in the past. For on-line feedback estimation of the crucial variances V and $\mathbf{W}$ to be a feasible proposition - one which is reducible to a failsafe black-box diagnostic technique that requires the minimum of monitoring - it is necessary to consider fully the biases in the estimates of the forecast distribution, in relation to their size and implications on the ensuing feedback estimates $\widehat{V}$ and $\widehat{\mathbf{W}}$. In section 4.2, an exact algebraic form for the bias in the k^{th} step-ahead forecast error (resulting from misspecification of V and $\mathbf{W}$) is calculated in theorem 4.3, via two lemmata; a result not even addressed by previous authors advocating

the use of the forecast distribution in this area. From theorem 4.3, the inflation in the variance of the k^{th} step-ahead forecast error in the first-order polynomial (steady) model is calculated, in terms of the true and misspecified signal-to-noise ratios, $r_0 = W_0/V_0$ and $r = W/V$ respectively. Further, the inflation in the covariance between the first and second step-ahead forecasts is calculated, again in terms of the true and misspecified signal-to-noise ratios. Hence, given sample estimates of the variance of the first step-ahead forecast error, and the covariance between the first and second step-ahead forecast errors, it is possible to solve, very straightforwardly, for both the true observational variance V_0, and the true state evolution variance W_0.

Given *precise* estimates of the step-ahead forecast distribution, therefore, the feedback estimates of V_0 and W_0 would be exact also. Thus the only further considerations to be made relate to the convergence properties of these estimates, which are addressed next in section 4.3, through calculation of the variances of the both the aforementioned variance and covariance estimates in the steady model. If the specified signal-to-noise ratio r is overestimated (i.e. W is too large with respect to V), these variances remain finite and relatively small, so convergence of the relevant variance and covariance estimates is fast, resulting in rapid convergence of $\widehat{V}$ and $\widehat{\mathbf{W}}$. On the other hand, underestimation of r leads to slower convergence of the observational and state evolution variances. These results are then supported through extensive simulations, by forecasting a simulated steady model with various misspecified values of r which illustrate the considerable effectiveness of the entire approach.

These methods are tested again in a 'head-to-head' comparison with the

multi-process, sub-optimal filtering techniques of Cantarelis and Johnston. The comparison is made using Cantarelis and Johnston's own choice of illustration for their methodology, where they fit eight steady models, all with differing r specifications, to a chemical process series taken from Box and Jenkins (who themselves fitted an ARIMA(0,1,1) process to this series). The results show the rapid convergence of the exact feedback approach compared to Cantarelis and Johnston's more laborious and time-consuming method, although we are of course no longer able to compare estimates of V and W directly back to the 'correct' specifications.

Section 4.4 then progresses onto studying the application of the approach to the 2-dimensional TSDLM, a model class which is widely-used but for which few authors have previously attempted on-line variance estimation of any kind. Similar, but more involved, calculations are made, once again utilising theorem 4.3, with respect to the biases in the forecast distribution for the variances of the first and second step-ahead forecast errors, and the covariances between these two forecast errors (three identities are now needed, to solve for V together with both diagonal elements W_1 and W_2 of the 2-dimensional $\mathbf{W}$). This leads to three simultaneous quadratic equations in three unknowns, each of which is, in turn, a function of the three variables V, W_1 and W_2, so that again, if the estimates of the required forecast distribution terms were precise, it would be possible to solve for the true variances V_0 and $\mathbf{W}_0$ exactly.

The convergence properties of the relevant forecast distribution identities are too complex to attempt to derive for the 2-dimensional TSDLM, however, and so simulation results are presented directly, in which misspecified

2-dimensional TSDLMs are fitted to a simulated series. The results indicate both that the preceding calculations are correct, and that further there are some complications due to the necessary numerical solution of the three simultaneous quadratic equations; overall they illustrate the practical applicability of the entire methodology even to this more complicated model, leaving the concluding impression that the problem of variance estimation in the widely-used model class of constant TSDLMs has, for the first time, been solved *without* reference to sub-optimal filtering and its associated difficulties.

Related debate

In section 2.2, I mention that the Dynamic Linear Model has not been adopted as widely as expected in time series analysis after Harrison and Stevens' paper on Bayesian Forecasting, an issue first raised in Ameen and Harrison's paper on discounting in 1985, but still as poignant today. There are undoubtedly many practical reasons for this, some of which have been motivation for most of this thesis, but additionally many deep-rooted philosophical issues lie behind this resistance to change.

There is no doubting the brilliance of Harrison and Stevens' paper, especially its introduction on the ideals and aims of forecasting. I would do that particular passage an injustice to summarise it - I have listed the problems of the Box-Jenkins approach in the light of these forecasting aims already in chapter 2 - but would certainly wish to second G.J.A. Stern's view (in the discussions following the paper) that this section "ought to be set up as a permanent block

of type and incorporated in all subsequent papers and books on forecasting...". The clarity of this introduction is in evidence throughout the rest of the paper, which goes on to outline the DLM system, and further illustrate not only how all Box-Jenkins processes are contained within the DLM framework, but also how all the desirable properties of forecasting listed in that introduction are self-evident from following this Bayesian approach. So why, then, the slowness to catch on?

The key to this also lies in the discussions that follow the paper. In these comments, O.D. Anderson draws a neat summary of how the Harrison-Stevens' (HS) approach must be viewed in order to outstrip Box-Jenkins (BJ) of the frontrunner's position in time series analysis: "...in the end, it is Jack [you or me] who has to be won over, by an approach which works for *him*...". And the problem here is that HS is an undoubtedly difficult *system* to use, whereas BJ is a black-box *method* that requires little thought - at least within the context of one analysis as distinct from another - and can be readily applied upon methodically following a set of simple instructions. I use the same distinction here between a forecasting method and system as Harrison and Stevens do; the former implies exactly such a black-box 'input-output' routine, the latter implies interaction between practitioners, forecasts, and resultant decisions. And these interactions require careful consideration of all influencing factors on our data - not only in which model we choose to represent our data response with (how *does* the series evolve, and how do we parametrise it?), but in the choice of initial priors (what is my initial state, and how certain am I of this?), choice of observational and state evolution variances (to which chapters 3 and 4 are

devoted!), and also how and when to intervene in the light of external information received (I know that next week is Christmas week, so what is likely to happen to the mean level and daily variation of my bread-sales data? (see section 2.1)).

This is, of course, no criticism of the HS system; it is just the freedom we seek, in fact. But it is not only truly daunting for an inexperienced practitioner to be left so freely with the reins, it is also expensive (both in terms of training time and, potentially, in terms of mistakes made whilst learning) to retrain a team to use a new system. "Certainly there will be effectively a step back, before two steps forward can again be taken", as Anderson puts it. Indeed, Harrison and Stevens state that the development of their system was only possible due to the generously ambitious support of ICI.

ICI management may not have been actively converted from BJ to HS, since the two approaches were developed almost in parallel, but the convertion of other Jacks who utilise BJ in the practitioners' world will only tend to follow as their financial decisions dictate - if BJ 'fails' for them in a particular crisis, the search for an improvement may lead to HS. This motivation is indeed a purely financial one. But we must remember that there is another motivation - that from the academic world.

When the inexperienced Jack (someone who has not yet been 'won over' by any particular approach) is searching for new pastures to solve an original forecasting problem, his conditioned reflex is almost always to be moved to seek what everyone else would use under the same conditions. This search in turn develops via word of mouth, if he is lucky enough to be in contact with a Mr.

Jones-next-door who has had a similar trouble, or, more often, via a literature review of theoretical and applied statistical publications. And this is where the leading lights of this world fail in their communication with Jack at the first hurdle. Harrison and Stevens' paper originally contained several examples illustrating the usage of the DLM, but these were cut from the final paper to reduce it to 'conventional length', with the result that a large majority of subsequent criticisms in the discussion stemmed from a lack of comprehension of how to apply the HS system. Yet it is vital that the application of such a *system* is illustrated, for as already mentioned, each analysis will be dealt with uniquely within the context of the problem. Accordingly, we find such remarkable comments in the discussion as, from Chatfield:

> *"One problem the paper leaves unanswered is the identification problem. The authors show that their DLM contains nearly every other forecasting procedure but we are not told how to select the appropriate model for a particular case. Perhaps their model is* too *general."*

Here we have the statistical academic world looking for a black-box method again, in a system which had been designed especially to leave many of these methods behind. If Jack reads such a misrepresentation (which he surely will, being a conscientious sort), he is bound to close the book forever and look elsewhere! As it is, Jack will be looking for examples - illustrations of genuine hands-on applications - to guide him in his convertion, and lo-and-behold the Society axes these as being unimportant...

This is an exceptionally common issue, and one which restrains the widespread

adoption of Bayesian statistics in far more general terms than simply the DLM. It is succinctly expressed by Wooff, in his discussion on the opening address by Lindley in Bayesian Statistics 4:

"Firstly, many of us are academics, subject to the market forces operating on ivory towers: we must produce streams of papers to survive - quantity is essential, quality a welcome bonus. Where we suffer in comparison to frequentist statisticians is that we produce hard but meaningful analyses rather than easy, but arguably worthless analyses. In short, the easiest way for a Bayesian to publish is to publish theory, rather than to go to the trouble of performing and reporting Bayesian applications. In this way we dig our own graves: we simply cannot convince users of statistical methodologies of the efficacy of the Bayesian way without adding meat to our theoretical bones in the way of large numbers of successful applications.

The tilting of the balance towards theory rather than application is compounded by a belief that to do the former is somehow smarter. We must avoid the folly of the theoretician sneering at the practitioner: we should instead condemn the minimal stature of the theoretician divorced from reality."

(Incidentally, this makes Lindley's own discussion of the Harrison and Stevens' paper - his criticism of ICI management, for letting two "such able persons leave their staff", and then making the comment that "perhaps academia is the only place for creativity" - even more surprising, since it also somehow fails to recognise the significance of the two leading pieces to be written on forecasting up to that time (by Box and Jenkins, and Harrison and Stevens) having had their roots entirely within industry...)

At the same time, it is equally important that the practitioners do not sneer at the theoreticians. This is often motivated by a lack of technical understanding, and is the source of that most irritating of comments, said to me up to now by taxi cab driver and fellow graduate student alike (and which Wooff

comes perilously close to saying too): "wait till you enter the *real* world" [1]. The essence of successful (and I use the word advisedly) research must be that theory and practice go hand-in-hand, so that the reader is led from full statement of his starting problem (so that he knows he is holding the correct hand), through technical justification and resolution of the issues concerned, and then into full illustration of the solution, and not left stranded at any of these three equally important gates. Harrison and Stevens' paper is greatly devalued by stranding the reader at the third, illustrative gate, and likewise the sister paper of examples will be a hollow piece, having lost the theoretical justification that is at its heart.

And so we return to the motivation behind the contents of this thesis given earlier. The best way to convince Jack to hop over the fence into your back garden is actually to cross over to him first, and then lead him all the way back yourself, before showing him exactly what he has achieved on his journey. Throughout what follows, I have always endeavoured to motivate the reader through problematic examples, and hence give a full statement of the need for improvement; this is, in turn, followed by (generally) complete calculations and solutions of these problems. However, I have often wished for more original data to be available through the course of this research, which would have allowed more steps to be taken at the vital third stage without relying so heavily on simulations for illustration of methodology. Maybe this is merely a reflection on the isolation of the Ph.D. student from the 'real world' after all...

[1] Most irritating, that is, after "lies, damn lies and statistics".

Chapter 1

The Portmanteau Statistic

1.1 Historical overview of Box-Jenkins time series analysis

Since many of the fundamental aims and principles of time series analysis will be covered later, at the beginning of chapter 2, all that remains of relevance to be drawn here is a brief picture of the historical development of the first classic account in the field - G.E.P. Box and G.M. Jenkins' *Time Series Analysis, Forecasting and Control* [3]. Modelling dependence in time series had at last become a feasible possibility around 1960, with the advent of computing power capable of dealing with both large data sets and the enormity of calculating (for instance) an autocorrelation function (acf) or a partial acf. Many specific approaches were developed, such as Holt-Winters' [20] & [38], and Brown's [6], exponentially weighted moving averages, but Box-Jenkins developed an entire method of analysis, fully defining not only model representation from a well-

defined family of processes, but identification of the 'best' such process to fit the data available thus far, diagnostic checking of the adequacy of this model fit, and then forecasting initiation.

This entire black-box method was truly a classic standardised input-output routine for the practitioners' world, and as such has been both widely used and adapted since its appearance in 1970, to an extent where there are few improvements to be made other than in its major conceptual foundations. However, these alter the outlook of the statistician so radically that a whole new approach - system, even - is required; one such alternative comes from a Bayesian viewpoint and is introduced in chapter 2. The only area of the black-box that is not firmly set in concrete foundations is the second stage of the analysis, the diagnostic procedures during model fitting.

1.2 The Box-Jenkins family of models

Before we direct attention to this area, we must establish some notation and terminology. It is presumed that the reader will be more than familiar with Box-Jenkins time series analysis in general; very briefly, the basic class of models within the method is the well-known ARMA(p, q) stochastic process of

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t = (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q) \epsilon_t \ ,$$

where $\{X_t\}$ is the time series sequence of (generally discrete) observations, $\{\epsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma_\epsilon^2)$ white noise, the polynomial $\phi(B)$ is the autoregressive operator of order p, $\theta(B)$ is the

moving average operator of order q, and B is the backward shift operator (such that $BX_t = X_{t-1}$). There are familiar conditions on these AR(p) and MA(q) processes to ensure stationarity, and we can endeavour to induce stationarity on the series by differencing it d times, producing the more general ARIMA(p, d, q) (I for Integrated) process.

Once the appropriate process has been identified, through analysis of the acf and partial acf of the data series $\{X_t\}$, we can fit the model to the original series and go on to calculate the estimated residuals $\{\hat{\epsilon}_t\}$ of the process. It is the distribution of these that we largely study in any subsequent diagnostic check of "goodness of fit" of the model; one such diagnostic statistic is the portmanteau test, so-called as it has a standard distribution easily applicable to any fitted model.

1.3 Background to the portmanteau test

An important tool for model fitting in any form of time series analysis is the residual acf, given by

$$r_k = \frac{\sum_{t=k+1}^{n} \epsilon_t \epsilon_{t-k}}{\sum_{t=1}^{n} \epsilon_t^2}$$

for a time series of length n.

Various diagnostic techniques for the checking of a fitted model involve examining the distribution of the residual acf. Since it is readily shown that $\mathrm{Var}(r_k) = (n-k)/n(n+2) \simeq 1/n$ for k small relative to n, and as these r_k's are assumed to be independent and asymptotically Normal, we are naturally

led into considering one of these potential diagnostics,

$$S(r) = n \sum_{k=1}^{m} r_k^2 \sim \chi_m^2 \ .$$

However, this statistic is of little practical interest, since we only have available the *estimated* residual acf $\hat{r}_k$ arising from the fitted model, given as

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} \hat{\epsilon}_t \hat{\epsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\epsilon}_t^2} \ ,$$

and it was shown by Durbin [11] in 1970 that, unfortunately, the χ_m^2 distribution is not applicable to $S(\hat{r})$.

Box and Pierce [4] continue the discussion, though, and look at the *actual* variances of the estimated residual acf in relation to the AR(p) process. By a linear expansion of $\hat{\mathbf{r}}$ about $\mathbf{r}$, their result is to show that

$$\hat{\mathbf{r}} = (I - Q)\mathbf{r} \ , \tag{1.1}$$

for I equal to the $m * m$ identity matrix, and Q given by $Q = X(X^T X)^{-1} X^T$, where

$$X = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \psi_1 & 1 & 0 & \cdots & \vdots \\ \psi_2 & \psi_1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \psi_1 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ \vdots & \vdots & \vdots & & \psi_1 \\ \vdots & \vdots & \vdots & & \vdots \\ \psi_{m-1} & \psi_{m-2} & \cdots & \cdots & \psi_{m-p} \end{pmatrix} .$$

(The ψ_k's are found by expressing the fitted process as an infinite MA process, namely $X_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}$, for $\psi_0 = 1$). By taking m large relative to n, so that $\psi_j \simeq 0$ for $j \geq m$, this result leads to the practically applicable and very well-known 'portmanteau test' of looking at

$$S(\hat{r}) = n \sum_{k=1}^{m} \hat{r}_k^2 \sim \chi^2_{m-p-q}$$

(for a fitted ARMA(p,q) model).

This was widely accepted and used until Pierce [33] in 1977 commented that the portmanteau test "needs more work", and this seemed to spark off a spate of research into its performance. Davies et al. [9] obtained expressions for the exact mean and variance of S, by dropping the usual assumption that the $\hat{r}_k$'s are Normally distributed. These calculated means and variances were compared with simulation results for fitting an AR(1) model for various sample sizes, all of which indicate that the statistic S seriously underestimates the true significance levels of lack of fit unless m is *small* relative to n (typically $n \geq 500$ for $m = 20$).

In a parallel report, Ljung and Box [26] also survey the accuracy of S, concluding that it gave suspiciously low values for its true distribution of χ^2_{m-p-q} (for a fitted ARMA(p,q) model), and suggesting the use of

$$S'(\hat{r}) = n(n+2) \sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k} \sim \chi^2_{m-p-q}$$

as a logically more accurate alternative test statistic. However, Davies et al. also

obtain an exact expression for the variance of S' when fitting an AR(1) model to white noise, and this indicates that despite the mean of S' being much closer to its asymptotic value of $m-1$, its variance for smaller samples could now be seriously inflated over the χ^2_{m-p} value of $2(m-1)$. Their attempt at numerically demonstrating the weaknesses of S involved the interesting approximation of this statistic's distribution by a central $a\chi^2_\nu$ density. Although exhibiting promising accuracy for an AR(1) model with smaller ϕ values, there were still disturbing inaccuracies for $\phi = 0.9$, caused by larger-than-expected tails in the actual density of S. In addition, this $a\chi^2_\nu$ approximation immediately renders the 'portmanteau' part of the test extremely inappropriate, as the distribution of the statistic would now be different for each model fitted! Indeed, the conclusion in that paper was that "rather less faith should be put in the portmanteau test....".

It was about now that the statistical world appeared to accept that the portmanteau test could never be anything more than just a poor guide for goodness-of-fit in diagnostic checking, and decided to avoid its problems by looking for alternative tests, notably extensions into multivariate AR models. Despite many attempts, alternative statistics were either shown to be equivalent to the portmanteau test under certain circumstances (Godfrey, [14], 1979; Hosking, [21], 1980 and the Lagrange multiplier test; Poskitt and Tremayne, [34], 1981 and the 'score test statistic'), or else made gains in areas not related to accuracy of performance (Godolphin, [15], 1980)). Throughout all of this, it is poignant to read that most applied statisticians (notably the hydrologists and economists) were quite content to continue quoting the portmanteau test

as the last word in goodness-of-fit testing. Indeed, some fifteen years on from Davies et al.'s work, there are still many areas in which the application of the portmanteau test is common. Not least is its use in the statistical package S [35], which still quotes significance values based upon the use not even of the Ljung-Box statistic S', but of Box-Pierce's S!

The sparseness of research into *why* the portmanteau test can give such inaccurate results is evident from a comment by Milhoj [32] in 1981, who derived a frequency domain analogue of the portmanteau statistic. In this paper, Milhoj wrote that

> *"in practical use it is often noted that the Box-Pierce portmanteau test, or the modified test, is weak, but no theoretical work has been done to explain why".*

Through further calculations of exact *general* expressions for the mean and variance of S', we find precisely why this latter quantity is indeed substantially inflated when the sample size n is small. There is a real need to improve its performance in these cases, since these are the very conditions under which it is most commonly practically applied.

1.4 Theory pertaining to the distribution of S'

From equation 1.1 we have that

$$\begin{aligned}\sum_{k=1}^{m} \hat{r}_k^2 = \hat{\mathbf{r}}^T\hat{\mathbf{r}} &= \mathbf{r}^T(I-Q)^T(I-Q)\mathbf{r} \\ &= \mathbf{r}^T(I-Q)\mathbf{r} \quad \text{(since } Q \text{ is idempotent)} .\end{aligned}$$

Then from the definition of S we can readily continue to derive exact expressions for $\mathrm{E}[S]$ and $\mathrm{Var}[S]$, which no longer involve the assumption of Normality of the $\hat{r}_k$'s. However, the definition of S' includes the $(n-k)^{-1}$ scale factor, and this rather complicates the picture. Defining an $m \times m$ diagonal matrix

$$D = \begin{pmatrix} (n-1)^{-1} & & 0 \\ & \ddots & \\ 0 & & (n-m)^{-1} \end{pmatrix}$$

then gives

$$\begin{aligned} \frac{S'}{(n(n+2))} = \sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k} &= \hat{\mathbf{r}}^T D \hat{\mathbf{r}} \\ &= \mathbf{r}^T (I-Q)^T D (I-Q) \mathbf{r} \ . \end{aligned}$$

So defining $A = D - DQ - QD + QDQ$ gives

$$\frac{S'}{(n(n+2))} = \mathbf{r}^T A \mathbf{r} = \sum_{k=1}^{m} A_{kk} r_k^2 + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} A_{jk} r_j r_k \ , \qquad (1.2)$$

(for A_{jk} equal to the $(j,k)^{th}$ element of A).

$$\begin{aligned} \text{Hence} \quad E[S'] &= n(n+2) \sum_{k=1}^{m} A_{kk} \mathrm{E}[r_k^2] \ \ (\text{since } \mathrm{E}[r_j r_k] = 0 \text{ for } j \neq k) \\ &= \sum_{k=1}^{m} A_{kk}(n-k) \ \ (\text{from } \mathrm{Var}(r_k) = (n-k)/(n(n+2)) \ ) \ , \end{aligned}$$

and so the expected value of the square of equation 1.2 minus the square of $\mathrm{E}[S']$ gives

$$\mathrm{Var}(S') = n^2(n+2)^2 \sum_{k=1}^{m} A_{kk}^2 \mathrm{Var}(r_k^2) + 2n^2(n+2)^2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} A_{kk} A_{jj} \mathrm{Cov}(r_k^2, r_j^2)$$

$$+ \; 4n^2(n+2)^2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} A_{jk}^2 \mathrm{E}[r_k^2 r_j^2] \; . \tag{1.3}$$

The first term alone in this expression would result from the usual asumptions of independence and asymptotic Normality of the r_k's. So we are now in a position to examine how the actual variance of S' is inflated above its asymptotic χ^2 value. For if we consider the AR(1) model, $(1-\phi B)X_t = \epsilon_t$, we easily find the ψ_k weights (in the definition of X above) to be given as ϕ^k, enabling us to calculate Q, leading to

$$(I-Q) = \begin{pmatrix} \phi^2 & -\phi+\phi^3 & -\phi^2+\phi^4 & \cdots \\ -\phi+\phi^3 & 1-\phi^2+\phi^4 & -\phi^3+\phi^5 & \cdots \\ \vdots & & \ddots & \end{pmatrix}.$$

We then readily proceed to calculating A as defined above. If we repeat these calculations for an AR(2) process, $(1-\phi_1 B-\phi_2 B^2)X_t = \epsilon_t$, the immediate problem comes in finding an expression for the ψ_k weights in the matrix X - these are found from the following theorem:

Theorem 1.1: When the AR(2) model, $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$, is expanded as the infinite MA model, $X_t = \sum_0^\infty \psi_k \epsilon_{t-k}$, $(\psi_0 = 1)$, we have that:

$$\psi_k = \phi_1^k + (k-1)\phi_1^{k-2}\phi_2 + \frac{(k-2)(k-3)}{2!}\phi_1^{k-4}\phi_2^2 + \ldots + \frac{(k-j)\ldots(k-2j+1)}{j!}\phi_1^{k-2j}\phi_2^j$$
$$+ \; \ldots + \left\{ \begin{array}{ll} \phi_2^{\frac{k}{2}} & \text{for } k \text{ even} \\ \frac{(k+1)}{2}\phi_1\phi_2^{\frac{(k-1)}{2}} & \text{for } k \text{ odd} \end{array} \right\}, \; k = 0, 1, \ldots \; .$$

Proof: Evidently, $\psi_0 = 1 = \phi_1^0$ and $\psi_1 = \phi_1^1$, obtained from the first substi-

tution into the AR(2) model above. Now assume that ψ_k has the form given for $k = 0, 1, \ldots, i$. Then $X_t = \sum_0^\infty \psi_k \epsilon_{t-k}$ in the AR(2) model form gives:

$$\begin{aligned} \sum \psi_k \epsilon_{t-k} &= \phi_1 \sum \psi_k \epsilon_{t-k-1} + \phi_2 \sum \psi_k \epsilon_{t-k-2} + \epsilon_t \\ \Rightarrow \sum_1^\infty \psi_k \epsilon_{t-k} &= \phi_1 \sum_0^\infty \psi_k \epsilon_{t-k-1} + \phi_2 \sum_0^\infty \psi_k \epsilon_{t-k-2} \end{aligned}$$

Considering coefficients of ϵ_{t-i-1} gives:

$$\psi_{i+1} = \phi_1 \psi_i + \phi_2 \psi_{i-1} \quad .$$

Hence from our inductive assumption above, we can consider the RHS of this last expression:

$$\phi_1 \psi_i = \phi_1^{i+1} + (i-1)\phi_1^{i-1}\phi_2 + \frac{(i-2)(i-3)}{2!}\phi_1^{i-3}\phi_2^2 + \frac{(i-3)(i-4)(i-5)}{3!}\phi_1^{i-5}\phi_2^3$$
$$+ \ldots + \frac{(i-j)(i-j-1)\ldots(i-2j+1)}{j!}\phi_1^{i-2j+1}\phi_2^j + \ldots + \begin{cases} \phi_1 \phi_2^{\frac{i}{2}} & , \ i \text{ even} \\ \frac{(i+1)}{2}\phi_1^2 \phi_2^{\frac{(i-1)}{2}} & , \ i \text{ odd} \end{cases}$$

and

$$\phi_2 \psi_{i-1} = \phi_1^{i-1}\phi_2 + (i-2)\phi_1^{i-3}\phi_2^2 + \frac{(i-4)(i-3)}{2!}\phi_1^{i-5}\phi_2^3$$
$$+ \ldots + \frac{(i-j)(i-j-1)\ldots(i-2j+2)}{(j-1)!}\phi_1^{i-2j+1}\phi_2^j + \ldots + \begin{cases} \phi_2^{\frac{i+1}{2}} & , \ (i-1) \text{ even} \\ \frac{i}{2}\phi_1 \phi_2^{\frac{i}{2}} & , \ (i-1) \text{ odd} \end{cases}$$

Thus

$$\begin{aligned}
\text{RHS} &= \phi_1^{i+1} + i\phi_1^{i-1}\phi_2 + \left((i-2) + \frac{(i-2)(i-3)}{2!}\right)\phi_1^{i-3}\phi_2^2 \\
&+ \left(\frac{(i-3)(i-4)}{2!} + \frac{(i-3)(i-4)(i-5)}{3!}\right)\phi_1^{i-5}\phi_2^3 \\
+\ldots &+ \left(\frac{(i-j)\ldots(i-2j+1)}{j!} + \frac{(i-j)\ldots(i-2j+2)}{(j-1)!}\right)\phi_1^{i-2j+1}\phi_2^j \\
+\ldots &+ \begin{cases} \phi_1\phi_2^{\frac{i}{2}} + \frac{i}{2}\phi_1\phi_2^{\frac{i}{2}} \ , & i \text{ even} \\ \phi_2^{\frac{i+1}{2}} + \frac{i+1}{2}\phi_1^2\phi_2^{\frac{i-1}{2}} \ , & i \text{ odd} \end{cases}
\end{aligned}$$

$$\Rightarrow \text{RHS} = \phi_1^{i+1} + i\phi_1^{i-1}\phi_2 + (i-2)\left(\frac{(i-3)}{2} + \frac{2}{2}\right)\phi_1^{i-3}\phi_2^2 + \frac{(i-3)(i-4)}{2!}\left(\frac{(i-5)}{3} + \frac{3}{3}\right)\phi_1^{i-5}\phi_2^3$$

$$+\ldots + \frac{(i-j)\ldots(i-2j+2)}{(j-1)!}\left(\frac{(i-2j+1)}{j} + \frac{j}{j}\right)\phi_1^{i-2j+1}\phi_2^j$$

$$+\ldots + \begin{cases} \frac{(i+2)}{2}\phi_1\phi_2^{\frac{i}{2}} \ , & (i+1) \text{ odd} \\ \phi_2^{\frac{i+1}{2}} \ , & (i+1) \text{ even} \end{cases}$$

$$\Rightarrow \text{RHS} = \phi_1^{i+1} + i\phi_1^{i-1}\phi_2 + \frac{(i-1)(i-2)}{2!}\phi_1^{i-3}\phi_2^2 + \frac{(i-2)(i-3)(i-4)}{3!}\phi_1^{i-5}\phi_2^3$$

$$+\ldots + \frac{(i-j+1)(i-j)\ldots(i-2j+2)}{j!}\phi_1^{i-2j+1}\phi_2^j + \ldots + \begin{cases} \phi_2^{\frac{i+1}{2}} \ , & (i+1) \text{ even} \\ \frac{(i+2)}{2}\phi_1\phi_2^{\frac{i}{2}} \ , & (i+1) \text{ odd} \end{cases}$$

which is of the form given for ψ_k, $k = i+1$, in the theorem. This completes the proof by Induction.

It is both interesting and of use in subsequent calculations to see these ψ_k coefficients explicitly; here are the first few:

$$
\begin{array}{lllllll}
\psi_0: & 1 & & & & & \\
\psi_1: & \phi_1 & & & & & \\
\psi_2: & \phi_1^2 & + \quad \phi_2 & & & & \\
\psi_3: & \phi_1^3 & +2\phi_1\phi_2 & & & & \\
\psi_4: & \phi_1^4 & +3\phi_1^2\phi_2 & + \quad \phi_2^2 & & & \\
\psi_5: & \phi_1^5 & +4\phi_1^3\phi_2 & + 3\phi_1\phi_2^2 & & & \\
\psi_6: & \phi_1^6 & +5\phi_1^4\phi_2 & + 6\phi_1^2\phi_2^2 & + \quad \phi_2^3 & & \\
\psi_7: & \phi_1^7 & +6\phi_1^5\phi_2 & +10\phi_1^3\phi_2^2 & + 4\phi_1\phi_2^3 & & \\
\psi_8: & \phi_1^8 & +7\phi_1^6\phi_2 & +15\phi_1^4\phi_2^2 & +10\phi_1^2\phi_2^3 & + \quad \phi_2^4 & .
\end{array}
$$

Theorem 1.1 can now be used directly to calculate A longhand, but it is worth noting from the definition of X that

$$
(X^TX)^{-1} = \frac{1}{(\sum\psi_k^2)^2 - (\sum\psi_k\psi_{k-1})^2}\begin{pmatrix} \sum\psi_k^2 & -\sum\psi_k\psi_{k-1} \\ -\sum\psi_k\psi_{k-1} & \sum\psi_k^2 \end{pmatrix} .
$$

Then from expressing the AR(2) model as the infinite MA process, we find

$$
E[X_t^2] = \sigma_\epsilon^2 \sum_0^\infty \psi_k^2 \ , \text{ since } \mathrm{E}[\epsilon_{t-i}\epsilon_{t-j}] = \begin{cases} 0 & \text{for } i \neq j \\ \sigma_\epsilon^2 & \text{for } i = j \end{cases} .
$$

Hence $\mathrm{Var}(X_t) = \sigma_\epsilon^2 \sum_0^\infty \psi_k^2$, since $E[X_t] = 0$. But directly from the AR(2) model, we get

$$
\begin{aligned}
\mathrm{Var}(X_t) &= \frac{\sigma_\epsilon^2}{\left(1 - \frac{\phi_1^2}{1-\phi_2} - \phi_2\left(\frac{\phi_1^2}{1-\phi_2} + \phi_2\right)\right)} \\
&= \frac{\sigma_\epsilon^2(1-\phi_2^2)}{(1-\phi_2^2)^2 - (\phi_1(1+\phi_2))^2} \ , \\
\Rightarrow \sum_0^\infty \psi_k^2 &= \frac{(1-\phi_2^2)}{(1-\phi_2^2)^2 - (\phi_1(1+\phi_2))^2} .
\end{aligned}
$$

Similarly, we have

$$
E[X_tX_{t-1}] = \gamma_1 = \sigma_\epsilon^2 \sum \psi_k\psi_{k-1}
$$

(where γ_1 is the first lag autocovariance), and also (for ρ_1 equal to the first lag autocorrelation)

$$\begin{aligned}\gamma_1 &= \rho_1\gamma_0 = \frac{\phi_1}{1-\phi_2}\cdot\frac{\sigma_\epsilon^2(1-\phi_2^2)}{(1-\phi_2^2)^2-(\phi_1(1+\phi_2))^2} \\ &= \frac{\sigma_\epsilon^2\phi_1(1+\phi_2)}{(1-\phi_2^2)^2-(\phi_1(1+\phi_2))^2}\,.\end{aligned}$$

Hence

$$\sum\psi_k\psi_{k-1} = \frac{\phi_1(1+\phi_2)}{(1-\phi_2^2)^2-(\phi_1(1+\phi_2))^2}\,,$$

giving us

$$(\sum\psi_k^2)^2-(\sum\psi_k\psi_{k-1})^2 = \frac{1}{(1-\phi_2^2)^2-\phi_1^2(1+\phi_2)^2}\,.$$

Thus

$$(X^TX)^{-1} = \begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ \phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix},$$

eventually giving us

$$I-Q = \begin{pmatrix} \phi_2^2 & a\psi_1+b & \cdots & a\psi_{i+j-2}+b\psi_{i+j-3} & \cdots \\ a\psi_1+b & 1+a(1+\psi_1^2)+2b\psi_1 & \cdots & \cdots & \cdots \\ \vdots & & \ddots & & \\ a\psi_{i+j-2}+b\psi_{i+j-3} & \cdots & \cdots & 1+a(\psi_{i-1}^2+\psi_{i-2}^2)+2b\psi_{i-1}\psi_{i-2} & \cdots \\ \vdots & & & & \ddots \end{pmatrix}, \tag{1.4}$$

(where $a=(\phi_2^2-1)$ and $b=\phi_1(1+\phi_2)$). Calculation of A can now be completed; in addition, if we consider again the assumption of asymptotic Normality and

independence of $\hat{\mathbf{r}}$, what we have shown is that under these conditions

$$\mathrm{Var}(\hat{r}_k) = \left(1 + (\phi_2^2 - 1)(\psi_{k-1}^2 + \psi_{k-2}^2) + 2\phi_1(1+\phi_2)\psi_{k-1}\psi_{k-2}\right)\mathrm{Var}(r_k) \ , \quad (1.5)$$

which in particular gives

$$\left.\begin{aligned} \mathrm{Var}(\hat{r}_1) &= \phi_2^2 \tfrac{(n-1)}{n(n+2)} \\ \text{and} \quad \mathrm{Var}(\hat{r}_2) &= ((\phi_1(1+\phi_2))^2 + \phi_2^2)\tfrac{(n-2)}{n(n+2)} \end{aligned}\right\} \quad (1.6)$$

These latter two results are also derived in Box and Pierce; they reconfirm the care that must be taken when looking at the size of $\hat{r}_k$ for lower lags k as a diagnostic check, due to the potentially large deflation of their actual variances from those of the r_k's.

However, the general result 1.5, combined with the earlier evaluation of the ψ_j's, enable us to also look at $\mathrm{Var}(\hat{r}_k)$ for higher lags k. For $\phi_1 = 1.6$, $\phi_2 = -0.9$, these results give $\mathrm{Var}(\hat{r}_3) = 0.840\mathrm{Var}(r_3)$, $\mathrm{Var}(\hat{r}_4) = 0.841\mathrm{Var}(r_4)$, and even for k as large as 9, $\mathrm{Var}(\hat{r}_9) = 0.911\mathrm{Var}(r_9)$. Hence the return to asymptotic behaviour of the $\hat{r}_k$'s as k increases is not necessarily as rapid as indicated by Box and Pierce, and we must remain wary of using simply the size of residual acf's as a diagnostic tool even for much higher lags than simply $k = p$.

1.5 Evaluation of Var(S')

Our final evaluations of A lead to the first 6 rows and 5 columns of the matrix (remembering its symmetry) being as follows for various AR processes. For

$n = 50$:

AR(1), $\phi = 0.1$,

$$A = \begin{pmatrix} 2.08\text{x}10^{-4} & -2.06\text{x}10^{-3} & -2.11\text{x}10^{-4} & -2.15\text{x}10^{-5} & -2.20\text{x}10^{-6} \\ -2.06\text{x}10^{-3} & 2.06\text{x}10^{-2} & -2.15\text{x}10^{-5} & -2.19\text{x}10^{-6} & -2.24\text{x}10^{-7} \\ -2.11\text{x}10^{-4} & -2.15\text{x}10^{-5} & 2.13\text{x}10^{-2} & -2.24\text{x}10^{-7} & -2.29\text{x}10^{-8} \\ -2.15\text{x}10^{-5} & -2.19\text{x}10^{-6} & -2.24\text{x}10^{-7} & 2.17\text{x}10^{-2} & -2.33\text{x}10^{-9} \\ -2.20\text{x}10^{-6} & -2.24\text{x}10^{-7} & -2.29\text{x}10^{-8} & -2.33\text{x}10^{-9} & 2.22\text{x}10^{-2} \\ -2.25\text{x}10^{-7} & -2.29\text{x}10^{-8} & -2.34\text{x}10^{-9} & -2.38\text{x}10^{-10} & -2.43\text{x}10^{-11} \end{pmatrix};$$

AR(1), $\phi = 0.9$,

$$A = \begin{pmatrix} 1.68\times10^{-2} & -3.28\times10^{-3} & -3.02\times10^{-3} & -2.78\times10^{-3} & -2.56\times10^{-3} \\ -3.28\times10^{-3} & 1.78\times10^{-2} & -2.77\times10^{-3} & -2.55\times10^{-3} & -2.35\times10^{-3} \\ -3.02\times10^{-3} & -2.77\times10^{-3} & 1.87\times10^{-2} & -2.35\times10^{-3} & -2.16\times10^{-3} \\ -2.78\times10^{-3} & -2.55\times10^{-3} & -2.35\times10^{-3} & 1.96\times10^{-2} & -1.99\times10^{-3} \\ -2.56\times10^{-3} & -2.35\times10^{-3} & -2.16\times10^{-3} & -1.99\times10^{-3} & 2.04\times10^{-2} \\ -2.36\times10^{-3} & -2.17\times10^{-3} & -1.99\times10^{-3} & -1.83\times10^{-3} & -1.68\times10^{-3} \end{pmatrix};$$

AR(2), $\phi_1 = 1.6,\ \phi_2 = -0.9$,

$$A = \begin{pmatrix} 1.66\times10^{-2} & -2.92\times10^{-3} & -1.18\times10^{-3} & 8.02\times10^{-4} & 2.43\times10^{-3} \\ -2.92\times10^{-3} & 1.75\times10^{-2} & -2.77\times10^{-3} & -1.37\times10^{-3} & 3.70\times10^{-4} \\ -1.18\times10^{-3} & -2.77\times10^{-3} & 1.79\times10^{-2} & -2.98\times10^{-3} & -1.67\times10^{-3} \\ 8.02\times10^{-4} & -1.37\times10^{-3} & -2.98\times10^{-3} & 1.81\times10^{-2} & -3.07\times10^{-3} \\ 2.43\times10^{-3} & 3.70\times10^{-4} & -1.67\times10^{-3} & -3.07\times10^{-3} & 1.88\times10^{-2} \\ 3.22\times10^{-3} & 1.89\times10^{-3} & 8.10\times10^{-5} & -1.64\times10^{-3} & -2.75\times10^{-3} \end{pmatrix}.$$

For $n = 500$:

AR(1), $\phi = 0.1$,

$$A = \begin{pmatrix} 2.01\times10^{-5} & -1.99\times10^{-4} & -1.99\times10^{-5} & -2.00\times10^{-6} & -2.00\times10^{-7} \\ -1.99\times10^{-4} & 1.99\times10^{-3} & -2.00\times10^{-6} & -2.00\times10^{-7} & -2.00\times10^{-8} \\ -1.99\times10^{-5} & -2.00\times10^{-6} & 2.01\times10^{-3} & -2.00\times10^{-8} & -2.01\times10^{-9} \\ -2.00\times10^{-6} & -2.00\times10^{-7} & -2.00\times10^{-8} & 2.02\times10^{-3} & -2.01\times10^{-10} \\ -2.00\times10^{-7} & -2.00\times10^{-8} & -2.01\times10^{-9} & -2.01\times10^{-10} & 2.02\times10^{-3} \\ -2.00\times10^{-8} & -2.01\times10^{-9} & -2.01\times10^{-10} & -2.02\times10^{-11} & -2.02\times10^{-12} \end{pmatrix};$$

AR(2), $\phi_1 = 1.6,\ \phi_2 = -0.9$,

$$A = \begin{pmatrix} 1.58\times10^{-3} & -3.16\times10^{-4} & -1.25\times10^{-4} & 8.45\times10^{-5} & 2.49\times10^{-4} \\ -3.16\times10^{-4} & 1.65\times10^{-3} & -2.93\times10^{-4} & -1.43\times10^{-4} & 3.46\times10^{-5} \\ -1.25\times10^{-4} & -2.93\times10^{-4} & 1.66\times10^{-3} & -3.06\times10^{-4} & -1.69\times10^{-4} \\ 8.45\times10^{-5} & -1.43\times10^{-4} & -3.06\times10^{-4} & 1.65\times10^{-3} & -3.02\times10^{-4} \\ 2.49\times10^{-4} & 3.46\times10^{-5} & -1.69\times10^{-4} & -3.02\times10^{-4} & 1.69\times10^{-3} \\ 3.22\times10^{-4} & 1.85\times10^{-4} & 5.35\times10^{-6} & -1.58\times10^{-4} & -2.59\times10^{-4} \end{pmatrix};$$

AR(2), $\phi_1 = \phi_2 = 0.1$,

$$A = \begin{pmatrix} 2.01\times10^{-5} & 2.21\times10^{-5} & -1.97\times10^{-4} & -1.75\times10^{-5} & -2.15\times10^{-5} \\ 2.21\times10^{-5} & 4.45\times10^{-5} & -1.95\times10^{-4} & -2.17\times10^{-4} & -4.12\times10^{-5} \\ -1.97\times10^{-4} & -1.95\times10^{-4} & 1.97\times10^{-3} & -2.35\times10^{-5} & -6.29\times10^{-6} \\ -1.75\times10^{-5} & -2.17\times10^{-4} & -2.35\times10^{-5} & 1.99\times10^{-3} & -4.77\times10^{-6} \\ -2.15\times10^{-5} & -4.12\times10^{-5} & -6.29\times10^{-6} & -4.77\times10^{-6} & 2.02\times10^{-3} \\ -3.92\times10^{-6} & -2.59\times10^{-5} & -2.99\times10^{-6} & -2.90\times10^{-6} & -5.90\times10^{-7} \end{pmatrix} .$$

To find $\mathrm{Cov}(r_k^2, r_j^2)$, we require the results

$$E[r_k^2 r_j^2] = \frac{(n-j)(n-k)+12(n-k)-8j}{n(n+2)(n+4)(n+6)} \quad (k>j)$$

$$\text{and} \quad E[r_k^2] = \mathrm{Var}(r_k^2) = (n-k)/n(n+2) .$$

Note that the covariance between r_j^2 and r_k^2 is independent both of the type of the model involved, and the parameters of that process. Indeed, we need only consider differing lengths of the series under analysis; for $n = 50$, the first 6 rows and 10 columns of this covariance matrix are given as (to 3 s.f.)

$$\mathrm{Cov}(r_k^2, r_j^2) = 10^{-5}\times \begin{pmatrix} 67.2 & 2.35 & 2.29 & 2.19 & 2.04 & 1.84 & 1.58 & 2.18 & 1.81 & 1.39 \\ 2.35 & 64.5 & 2.24 & 2.14 & 1.99 & 1.78 & 1.53 & 2.13 & 1.76 & 2.23 \\ 2.29 & 2.24 & 61.9 & 2.09 & 1.94 & 1.73 & 1.48 & 2.08 & 1.71 & 2.18 \\ 2.19 & 2.14 & 2.09 & 59.4 & 1.89 & 1.68 & 1.43 & 2.03 & 1.65 & 2.12 \\ 2.04 & 1.99 & 1.94 & 1.89 & 56.8 & 1.63 & 2.29 & 1.97 & 1.60 & 2.07 \\ 1.84 & 1.78 & 1.73 & 1.68 & 1.63 & 54.4 & 2.24 & 1.92 & 1.55 & 2.01 \end{pmatrix} ,$$

and for $n = 500$, the first 6 rows, 10 columns are

$$\mathrm{Cov}(r_k^2, r_j^2) = 10^{-8}\times \begin{pmatrix} 785 & 3.10 & 3.09 & 3.08 & 3.06 & 3.04 & 3.01 & 3.08 & 3.04 & 2.99 \\ 3.10 & 783 & 3.09 & 3.07 & 3.06 & 3.03 & 3.00 & 3.07 & 3.03 & 3.09 \\ 3.09 & 3.09 & 780 & 3.07 & 3.05 & 3.02 & 2.99 & 3.07 & 3.02 & 3.08 \\ 3.08 & 3.07 & 3.07 & 776 & 3.04 & 3.02 & 2.99 & 3.06 & 3.02 & 3.07 \\ 3.06 & 3.06 & 3.05 & 3.04 & 773 & 3.01 & 3.09 & 3.06 & 3.01 & 3.07 \\ 3.04 & 3.03 & 3.02 & 3.02 & 3.01 & 770 & 3.09 & 3.05 & 3.01 & 3.06 \end{pmatrix} .$$

The diagonal terms in both covariance matrices are found from knowing $\mathrm{E}[r_k^4]$, given in Davies et al., but in any case they are unnecessary for evaluation of

the second and third terms of $\mathrm{Var}(S')$ in equation 1.3.

The effects of these matrices on the inflation of $\mathrm{Var}(S')$ can now be studied. Firstly, note that in all of the above examples, the diagonal elements of A (the A_{kk} terms) actually increase in size, whereas the off-diagonal elements (the A_{jk} terms) decrease exponentially, due to the ϕ^{j+k} terms in $I - Q$. This decay in size is therefore very rapid when $|\phi|$ is small (in the AR(1) process) or $|\phi_1|$ and $|\phi_2|$ are small (for the AR(2)), but is somewhat slower for larger $|\phi|$, or ϕ_1 and ϕ_2 close to the stationarity boundaries. In either case the diagonal A_{kk}'s remain much larger than the off-diagonal A_{jk}'s, which renders the third term in equation 1.3,

$$4n^2(n+2)^2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} A_{jk}^2 \mathrm{E}[r_k^2 r_j^2] \ ,$$

negligible in comparison with the second term in 1.3,

$$2n^2(n+2)^2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} A_{kk} A_{jj} \mathrm{Cov}(r_k^2, r_j^2) \ . \tag{1.7}$$

Hence we need only concern ourselves with this latter expression, equation 1.7, when considering the source of the inflation of $\mathrm{Var}(S')$ above its asymptotic value of $2(m-p-q)$.

It is also worthy of note that the diagonal elements of A for $n = 50$ are approximately 10 times larger than the corresponding elements for $n = 500$. This is simply due to the the $1/(n-k)$ factor in the matrix D (it therefore also affects the off-diagonal elements of A similarly). If we perform a brief order-of-magnitude calculation, then taking the AR(1) model, with $\phi = 0.1$, we have for $n = 50$ and $m = 20$ that $2n^2(n+2)^2 A_{22} A_{33} \mathrm{Cov}(r_2^2, r_3^2) = 0.133$, and

as there are 190 terms in equation 1.7 with $m = 20$, with $\text{Cov}(r_k^2, r_j^2)$ decaying in size for increasing k, j, we should find an overall inflation in $\text{Var}(S')$ of somewhere around 20. If we repeat this for $n = 500$, the net effect of the $n^2(n+2)^2\text{Cov}(r_k^2, r_j^2)$ product is an increase by a factor of 10 - so when multiplied by A_{kk} and A_{jj}, both of which are decreased tenfold, we should find an overall reduction in size in the inflation of $\text{Var}(S')$ by a factor of 10, to around 2. This pattern is repeated for the other cases. Hence for small n there is apparently a quite serious inflation in $\text{Var}(S')$ (from 38 to about 58 in any AR(1) process), whereas for $n = 500$ the inflation is reasonably small.

The exact values for $\text{Var}(S')$ are calculated from equation 1.3, and shown for several AR(1) and AR(2) models in table 1.1 below, together with $\text{E}[S']$. In each case, $\text{Var}(S')$ is confirmed as being worryingly large for $n = 50$, and quite respectable for $n = 500$; $\text{E}[S']$ is always very close to its asymptotic value.

Table 1.1: Exact values of $\text{E}[S']$, $\text{Var}(S')$ (from 1.3) for various AR processes

						Exact values		χ^2 values	
p	ϕ	ϕ_1	ϕ_2	m	n	$\text{E}[S']$	$\text{Var}(S')$	$\text{E}[S']$	$\text{Var}(S')$
1	0.1			20	50	19	58.807	19	38
					500	19	40.501		
	0.5				50	19.0002	58.922		
					500	19	40.514		
	0.9				50	19.0094	59.097		
					500	19.0003	40.530		
2		0.1	0.1	20	50	18.0001	54.511	18	36
					500	18	38.231		
		1.0	-0.5		50	18.0016	54.777		
					500	18	38.262		
		-0.8	0.1		50	18.0107	54.812		
					500	18.0008	38.260		
		1.6	-0.9		50	18.0568	55.207		
					500	18.0318	38.294		

1.6 Improvement of S'

Returning to the evaluation of $I-Q$ for an AR(1) process in section 1.4, and ignoring the $\mathrm{Var}(r_k) \simeq \frac{1}{n}$ factor since it is lost in the later correlation calculations, we see that for this model

$$\mathrm{Cov}(\hat{r}_1, \hat{r}_2) = \phi^3 - \phi \ .$$

From equation 1.4 (the expression for $I-Q$ for an AR(2) process), by substituting the value $\psi_1 = \phi_1$ we find that

$$\begin{aligned} \mathrm{Cov}(\hat{r}_1, \hat{r}_2) &= (\phi_2^2 - 1)\phi_1 + \phi_1(1+\phi_2) \\ &= \phi_1\phi_2(1+\phi_2) \ . \end{aligned}$$

These two equations, together with expressions for $\mathrm{Var}(\hat{r}_1)$, $\mathrm{Var}(\hat{r}_2)$, lead to finding the correlation between $\hat{r}_1$ and $\hat{r}_2$ that follows from the assumption of Normality of $\hat{\mathbf{r}}$ in equation 1.1; we find

$$\rho(\hat{r}_1, \hat{r}_2) = \frac{-\phi}{|\phi|} \frac{1-\phi^2}{\sqrt{1-\phi^2+\phi^4}} \quad \text{for an AR(1)} \ ,$$

$$\text{and} \quad \rho(\hat{r}_1, \hat{r}_2) = \frac{\phi_1\phi_2(1+\phi_2)}{|\phi_2|\sqrt{\phi_1^2(1+\phi_2)^2+\phi_2^2}} \quad \text{for an AR(2)} \ .$$

In both models these correlations will, in certain cases, be large; either for ϕ small in an AR(1), or for $\phi_1 \to 1$, $\phi_2 \to 0$ (when $\rho(\hat{r}_1, \hat{r}_2) \to 1$) and $\phi_1, \phi_2 \to 0$ (when $\rho(\hat{r}_1, \hat{r}_2) \to 1/\sqrt{2}$) in an AR(2). It is also apparent from this Normality

of $\hat{\mathbf{r}}$ assumption that $|\mathrm{Cov}(\hat{r}_j, \hat{r}_k)|$ will be largest for $j = 1,\ k = 2$, and decrease exponentially as j, k increase, for an AR(1). In an AR(2) model, this decay in the covariances between the residual autocorrelations is more complicated, following a damped sinusoidal variation, but in all such processes the largest covariances are those involving $\hat{r}_1$ and $\hat{r}_2$.

When we come to examining the *actual* behaviour of S', however, it is $\mathrm{Cov}(r_k^2, r_j^2)$ that is of interest. But, as can be seen from the above matrices for these covariances, the same patterns are in evidence here. There is a damped sinusoidal variation by row or column, and for both $n = 50$ and $n = 500$ the largest value is $\mathrm{Cov}(r_1^2, r_2^2)$, with almost all the larger covariances coming in rows 1 and 2 - those involving r_1^2 and r_2^2.

The original aim of this exploration was to improve the performance of S' for small sample sizes. Under these conditions, we have seen that the statistic has a marginally inflated mean, but a grossly over-inflated variance. To succeed in our aim we must therefore decrease the variance of S' where possible. What we have also discovered, through studying their variances (in equation 1.6 earlier) and covariances, is that the first residual autocorrelation in an AR(1) model, and the first two, $\hat{r}_1$ and $\hat{r}_2$, in an AR(2) process, mostly have distributions far removed from the i.i.d. $N(0, \frac{n-k}{n(n+2)})$ assumed for the asymptotic χ^2 distribution of S' to hold. Hence the proposed alternative is to neglect the first p terms in S' when fitting an AR(p) model: to look at the alternative test statistic

$$S'' = n(n+2) \sum_{k=p+1}^{m} \frac{\hat{r}_k^2}{n-k} \sim \chi^2_{m-p} \,.$$

Following the above calculations for S', we achieve the results

$$\mathrm{E}[S''] = \sum_{k=p+1}^{m} A_{kk}(n-k)$$

$$\text{and } \mathrm{Var}(S'') = n^2(n+2)^2 \sum_{k=p+1}^{m} A_{kk}^2 \mathrm{Var}(r_k^2) + 2n^2(n+2)^2 \sum_{j=p}^{m-1} \sum_{k=j+1}^{m} A_{kk}A_{jj}\mathrm{Cov}(r_j^2, r_k^2)$$

$$+ \; 4n^2(n+2)^2 \sum_{j=p}^{m-1} \sum_{k=j+1}^{m} A_{jk}^2 \mathrm{E}[r_j^2 r_k^2] \; .$$

We can study the behaviour of these quantities for small sample sizes, ($n = 50$), with reference to the above evaluations of the matrices A and $\mathrm{Cov}(r_j^2, r_k^2)$. With an AR(1) model, $\mathrm{E}[S'']$ is decreased more for larger $|\phi|$, since A_{11} is greatly increased as $|\phi|$ increases (this is due to the ϕ^2 term as the first diagonal element of $I - Q$). $\mathrm{Var}(S'')$ is similarly affected, partly because of the increasing A_{11} term, but also because we are now ignoring the largest covariance terms in the variance inflating expression 1.7 - those involving $\hat{r}_1$. A very similar pattern is observed in relation to the AR(2) process, with A_{11} and A_{22} being much smaller than other diagonal elements in A when $|\phi_1|$, $|\phi_2|$ are small, but of similar magnitude for ϕ_1, ϕ_2 closer to the stationarity boundaries (again due to the exponential decay within the diagonal elements of $I - Q$ (see equation 1.4 earlier)).

All of these results are given below in table 1.2, together with the previous values from table 1.1 for S' as comparison.

Table 1.2: Exact values for means and variances of S', S'' for various AR models

p	ϕ	ϕ_1	ϕ_2	m	n	E[S']	E[S'']	Var(S')	Var(S'')
1	0.1			20	50	19	18.9898	58.807	58.738
					500	19	18.9900	40.501	40.458
	0.5				50	19.0002	18.7449	58.922	57.349
					500	19	18.7495	40.514	39.559
	0.9				50	19.0094	18.1838	59.097	55.186
					500	19.0003	18.1916	40.530	38.375
2		0.1	0.1	20	50	18.0001	17.9668	54.511	54.290
					500	18	17.9678	38.231	38.094
		1.0	-0.5		50	18.0016	17.2349	54.777	50.757
					500	18	17.2485	38.262	35.919
		-0.8	0.1		50	18.0107	17.2007	54.812	51.029
					500	18.0008	17.2102	38.260	36.135
		1.6	-0.9		50	18.0568	16.4056	55.207	47.766
					500	18.0318	16.4220	38.294	34.147

Note that for $n = 500$, the means and variances of S' are close enough to the asymptotic values to render S'' practically redundant - for smaller ϕ in an AR(1) there is little difference between the two statistics, as is the case for the AR(2) model with $\phi_1 = \phi_2 = 0.1$. For $\phi = 0.9$ in the AR(1), and in the other AR(2) models, S'' develops a more serious deflation of the mean (and variance, in the last model).

However, in all the $n = 50$ examples, the inflation in Var(S'') is indeed reduced as anticipated, with the added advantage that the mean is also deflated, to a varying degree. This will result in the improved accuracy of the test statistic S'' over S' when drawn from the χ^2_{m-p} distribution, for *all* fitted models on smaller samples.

1.7 Simulation results

Following the work of Box and Pierce [4], who showed the equivalence of the residual acf distribution from a correctly identified and fitted ARIMA(p,d,q) model with that from an ARIMA(p+q,d,0) process, it should be noted that all of the theoretical work to date in relation to an AR(p) process applies equally well to the behaviour of S, S' and S'' when dealing with either MA(q) or ARMA(p,q) models. Thus we can generalise our definition of the alternative statistic S'' to one of looking at

$$S'' = n(n+2) \sum_{k=p+q+1}^{m} \frac{\hat{r}_k^2}{n-k} \sim \chi^2_{m-p-q} \ ,$$

when fitting an ARMA(p,q) (or, of course, an ARIMA(p,d,q)) model.

The similarities within each statistic's distributions for differing model fits of the same order are evident in table 1.3 below, which gives the significance levels of S, S' and S'' at the 0.05 level when fitting an ARMA(p,q) model (for p+q$\leq$2) to a simulated process of the same identity of length n. The significance levels were calculated by simulating 1000 series for each different process, and then for each of the three statistics the total number of these series that produced a significant test result was taken as a ratio; (i.e. the first figure under the S' column implies that 68 of the 1000 simulated series for the AR(1) model $X_t - 0.1X_{t-1} = \epsilon_t$ gave a test value for S' that was significant at the 5% level).

Table 1.3: Significance levels (0.05 level) for the three Portmanteau statistics S, S' and S'', for fitting ARMA(p,q) processes to simulated series of the same type

								Significance levels			Means			Variances		
p	q	n	m	ϕ_1	ϕ_2	θ_1	θ_2	S	S'	S''	S	S'	S''	S	S'	S''
1	0	50	20	0.1				0.010	0.068	0.062	13.98	18.55	18.47	27.3	46.5	46.2
				0.5				0.019	0.074	0.069	14.24	18.92	18.61	31.9	56.0	55.2
				0.9				0.017	0.091	0.072	14.92	19.74	18.86	31.4	54.3	51.1
		500	20	0.1				0.046	0.054	0.054	18.46	18.95	18.93	37.2	39.2	39.1
				0.5				0.039	0.049	0.047	18.29	18.78	18.51	35.9	38.0	37.1
				0.9				0.042	0.053	0.040	18.63	19.12	18.32	37.6	39.6	37.7
			40	0.9				0.053	0.080	0.070	37.46	39.24	38.39	85.9	94.3	93.9
0	1	50	20			0.1		0.017	0.063	0.060	14.08	18.73	18.61	30.2	52.9	52.6
						0.9		0.019	0.093	0.078	14.94	19.69	18.86	32.4	55.2	51.8
		500	20			0.1		0.048	0.058	0.058	18.77	19.27	19.25	40.4	42.5	42.5
						0.9		0.042	0.045	0.039	18.78	19.26	18.48	35.3	37.2	35.6
2	0	50	20	0.1	0.1			0.006	0.054	0.047	12.61	16.98	16.78	22.3	39.2	38.7
		500	20					0.043	0.051	0.051	17.25	17.73	17.68	33.4	35.3	35.3
		50	20	1.0	-0.5			0.011	0.064	0.057	13.10	17.56	17.25	27.7	48.9	47.5
		500	20					0.042	0.052	0.036	17.34	17.81	17.00	36.4	38.4	35.4
		500	40					0.025	0.053	0.044	36.32	38.13	37.32	74.5	82.1	80.6
		50	20	-0.8	0.1			0.010	0.073	0.066	13.45	18.04	17.19	27.0	48.4	45.8
		500	20					0.050	0.055	0.044	17.73	18.22	17.46	36.6	38.6	37.3
		500	40					0.034	0.059	0.048	36.08	37.88	37.04	78.5	86.9	85.0
		50	20	1.6	-0.9			0.029	0.089	0.065	13.86	18.42	16.81	34.3	59.9	52.3
		500	20					0.046	0.052	0.030	17.95	18.42	16.75	33.8	35.5	31.5
		500	40					0.047	0.076	0.051	36.43	38.18	36.48	81.0	88.9	84.3
0	2	50	20			0.1	0.1	0.008	0.063	0.057	12.89	17.38	17.08	23.5	42.6	41.6
		500						0.041	0.052	0.052	17.39	17.87	17.82	35.0	37.0	36.8
		50	20			1.6	-0.9	0.017	0.065	0.045	13.34	17.66	15.96	29.3	50.1	42.4
		500						0.057	0.069	0.043	17.91	18.38	16.74	39.3	41.5	37.4
1	1	50	20	0.5		0.1		0.006	0.054	0.045	13.29	17.82	17.33	23.1	41.2	40.1
		500						0.040	0.043	0.040	17.33	17.81	17.48	32.7	34.5	33.6
		500	40					0.040	0.069	0.063	36.04	37.82	37.47	76.9	84.8	83.7
		50	20	-0.7		0.9		0.021	0.098	0.079	14.15	18.82	17.67	32.7	57.2	52.3
		500						0.037	0.043	0.032	17.60	18.08	17.21	34.3	36.2	33.8

It is important to remember that all three statistics have nothing more than approximate χ^2 distributions, and as such even S'' will have areas where its performance is also substandard. One is undoubtedly in large samples, where the inflation in the variance of S' is greatly reduced and so the significance levels of S'' (which has a yet smaller mean and variance) tend to underestimate

model inadequacy. However, it is interesting to note that as we increase m to 40 for a sample size of 500, the covariance sum in equation 1.7 above is greatly lengthened (from 190 terms to 780 in S'). Hence by testing up to lag 40 we increase the inflation in $\mathrm{Var}(S')$ once more well above its asymptotic value of $2(m-p-q)$. In nearly all of these cases, and especially so when the parameters of the process are close to the stationarity/invertibility boundaries, ignoring the first $p+q$ $\hat{r}_k$ terms then reduces this inflation (and deflates the means accordingly) to a level where S'' performs decidedly more accurately again.

The second area of potential concern in S'' is in the relative size of $p+q$ compared to the length of sum m. Clearly as $p+q$ increases with respect to m we will reach a point, even in small samples, where we are ignoring 'too many' terms at the beginning of the sum for the χ^2_{m-p-q} distribution to be a reasonable one still to assume for S''. Here, though, we are saved by the generally desirable statistical property of parsimony - to quote Box and Jenkins themselves: "In practice, it is frequently true that adequate representation of actually occurring stationary time series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than 2 and often less than 2" ([3], p.11). To perform further simulations for the ARMA(2,1), ARMA(1,2) and ARMA(2,2) processes would be not merely tedious, but unnecessary too, since it is evident from the $n = 50$ examples above that S'', whilst evidently being an improvement over S and S', still generally overestimates significances, even for $p+q = 2$ (in smaller sample sizes). Hence it would appear that the performance of S'' will be improved still further as we go on to consider $p+q = 3$ (or even 4, although these models are much less frequently employed) for small

sample sizes, given that the statistic is generally used for a summation length of at least $m = 20$.

1.8 Conclusions

These simulation results and observations lead us to two conclusions. Firstly they reconfirm the often disturbing inaccuracies in the significance levels of S for small samples, and confirm the suspicions of Davies et al. [9] that S' performs almost as badly in these situations (but now giving over-significant test results); for large samples, both statistics improve markedly in their performance. In fact, in the $n = 500$ examples, S' performs generally well enough to be quite acceptable. If, however, anyone should ever wish to test model fit up to higher lags, then S'' becomes more and more reliable as m increases.

The second and undeniable conclusion is that amongst all frequently fitted parsimonious ARMA(p,q) or ARIMA(p,d,q) processes, the test significance levels for small sample sizes are consistently improved, sometimes dramatically, when we look at S''.

Apart from showing how possible pitfalls can be encountered when studying simply the first few residual autocorrelations - unless due consideration is paid to their true distribution - clear illustration has also been made of the deliberation required throughout our use of the portmanteau test. Not only do we need great care in the conclusions that can be drawn from using this test, but also in our very *choice* of statistic, since account must be taken of both

the nature and length of the series under question. Accordingly, the portmanteau test should never be relied upon for black-or-white test results; like all diagnostic procedures, it is merely a potentially useful guide for possible model inaccuracies.

Chapter 2

Bayesian Dynamic Modelling

2.1 Introduction

The Box-Jenkins [3] approach to time series analysis was widely adopted and applied through the early 1970's. Series were analysed in sales forecasting and other econometric areas, hydrological applications, etc., etc. (see Fama & Schwert [12] and Mehta et al. [30], amongst many others, for illustration of time series analyses in these disciplines), and the applied statisticians of the world seemed relieved and grateful that at last they had an *ad hoc* mathematical method of describing their physical time series, and forecasting them accordingly. This relief, that the B.F.E. method (Bold Freehand Extrapolation, to borrow Chatfield and Prothero's [8] T.L.A.[1]) of forecasting was finally replaced by a black-box method, with concrete guidelines for model identification, application and goodness-of-fit testing, seemed to shelter practitioners from the underlying grave philosophical and practical worries with the whole approach.

[1]Three Letter Abbreviation

These are, under broad headings, fourfold:

(i) Conceptual interpretation of the ARIMA processes. In other words, what does the fitted model *mean*, physically, in the light of the series it is modelling?

(ii) The inability to cope quickly with - in terms of positive response - changes in the evolutionary nature of the series.

(iii) The frailty of the method in applications with little or no data. Box and Jenkins themselves state that "at least 50 or preferably 100 observations should be used" to enable 'correct' model identification.

(iv) The isolation of the practitioner and other sources of external information from the model.

These points are not in any particular order of statistical importance, but rather are listed in the order that they became apparent to the author in the course of studying time series analysis more broadly. During this study, the seeds of suspicion and doubt in the Box-Jenkins approach were slowly sown, and it was only a matter of time before sufficient evidence arose with respect to these worries to finally break the restraints - at least for this inexperienced practitioner - created by both stationarity, and the inability to express one's own knowledge about future events, that Box-Jenkins analysis is confined by.

The evidence can be as simple as this: suppose that you are interested in forecasting daily bread sales at a local bakery, recently opened early on in the summer (and so jolly keen to anticipate demand accurately), through the rest of the calendar year. Once you have enough data, you identify, say, a Box-Jenkins ARIMA(0,1,1) process (the parameters of which you update regularly as more data is gathered), together with a certain weekly seasonality. The model is both

intuitive - it is logically implied from a simple dynamic linear model that we meet later - and forecasting well. Then comes the middle of December, and on the night of Saturday the 17^{th} you have just forecasted Monday's demand (your baker chooses to stay shut on Sundays, allowing you time to consider your data), when the (really quite obvious) thought strikes you - the week up to Christmas will see much stocking-up for the two extra days of the following week that the bakery is shut, and the process will begin imminently. You now have an overwhelming feeling that your forecast for Monday is about $25-50\%$ too small, but as this is your first year of data, how do you adapt this figure? And as the seasonality for the week to come will almost certainly be altered, what do you do with regards to the rest of the week? Your model has become as redundant as you might well be by the end of that week!

In the 1970 version of "Time Series Analysis: Forecasting and Control", there is no answer to this problem in Box-Jenkins analysis. (Later in the 1970's, as the need for this interaction became more and more evident, methods for incorporating some form of intervention into the Box-Jenkins model were developed, but none of these methods were capapable of reflecting one's greatly increased uncertainties in the model at and beyond the point of intervention). It is crucial to realise that these are the very circumstances under which you *particularly* want your model to cope well, since such change points are evidently where forecasting performances are affected the most. Our bakery example is a perfect example of the points (ii) to (iv) made earlier. Not only does this illustrate how we are isolated from a model which, in turn, cannot respond to a change-point that we know is occurring, but our example (incidentally, despite

the dramatisation of the consultancy, the example does come from a 'real' data series, and is given as Series 1 in the Appendix) is also a failing of the Box-Jenkins approach in the face of little or no data. If we had come into the baker's lifetime at a point in several years' time, we could have not only anticipated this seasonal variation - for that is, of course, exactly how one would expect fluctuations to occur - but also built it into the Box-Jenkins model.

It is hoped that this example has frustrated even the disinterested reader at his or her complete helplessness in such a scenario. We are completely isolated from acting as an intellectual bridge between physical phenomena occurring in the environment that produce our data responses, and the mathematical model that we have built. But this bridging is, surely, a crucial role of the statistician!

Most of these issues were raised as early as 1973, in both Green and Harrison [16], who cite the absence of a yearly seasonality history of a marketed product (in our case, bread) as motivation for a Bayesian approach, and also in Chatfield and Prothero's [8] analysis of a carefully selected (as, indeed, the above bakery example was) sales series that was about to hit the recession of the next year, and so actually did have several years of seasonality to work with. Their at times wonderfully tongue-in-cheek paper was roundly criticised as not being fair to the Box-Jenkins method; however, it did highlight how careful the statistical world was being in only selecting series that were 'nicely behaved' to illustrate the powers of the approach. It is therefore not surprising that, albeit in more-or-less parallel with the development of Box-Jenkins analysis, the roots of a more general class of models were being nurtured in industry; a system capable of dealing with the Box-Jenkins shortcomings that were otherwise avoided. These

roots are in a Bayesian analysis, and are the subject of the rest of this thesis.

2.2 Development of the DLM

In 1976, in the Journal of the Royal Statistical Society, Series B, Professor P.J. Harrison and Mr. C.F. Stevens, an independent consultant, presented a paper simply entitled 'Bayesian Forecasting' [19]. Both men had been involved with ICI for many years previously, where the majority of their methods had been developed and applied. This broad wealth of industrial research and experience had produced many radical modelling and forecasting concepts, motivated largely by a desire to avoid the last three worries listed earlier - to be able to interact, adapt to structural disturbances in the model, and cope in periods of little or no data.

This paper was the birth - certainly in the statistical world, following their little-cited paper of 1971 in the Operational Research Quarterly [18] - of the Dynamic Linear Model (DLM) as an approach to time series analysis. The details will be defined later; simply, we can set up a model within the well-defined DLM class that represents the physical process of our system (making conceptual interpretation simple, of course, to return to the first worry of section 2.1), and then use the Kalman Filter (section 2.3), in the light of a new data point, to update *a priori* beliefs in this model and produce *a posteriori* beliefs. At each updating, we can easily build in changes in our prior beliefs and hence interact with the model; in addition, because of the very fact that we can represent our *beliefs* in the model, it can operate with little, or even no, data.

The power of the DLM is decidedly beyond question, but it is curiously interesting that nearly two decades later the DLM is not as widely adopted as might have been expected, both for philosophical and practical reasons, dwelt upon more in the Introduction and later in this chapter. It has, however, been developed and refined, and the next major leap forward came in 1985, in a paper by West, Harrison and Migon [37], on Dynamic Generalised Linear Models, which - as the title suggests - generalised the DLM to non-Normal models and applied them to, amongst others, the same data set as that used by Chatfield and Prothero [8] in 1973. The improvement in this forecasting test was evident, especially so - and this is the key consideration, of course - as observations from that next recessional year were processed. Together with several others, this paper formed the nucleus of a book, "Bayesian Forecasting and Dynamic Modelling", by West and Harrison [36] in 1989. In what follows, the notation of this book is adopted.

The basic univariate DLM (it is easy to generalise to multivariate data) consists of two defining equations. The first is the *observation* equation

$$Y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + v_t \ , \tag{2.1}$$

where Y_t is the univariate observation variable, $\boldsymbol{\theta}_t$ is the $p \times 1$ state vector, consisting of the p parameters of the defined process, $\mathbf{F}_t$ is a known $p \times 1$ vector of independent variables, and $v_t \sim \mathcal{N}(0, V_t)$ is the observational error. The

second equation is for the updating of the state vector, given by

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t \ , \tag{2.2}$$

and known as the *state* or *system* evolution equation, for the known $p \times p$ evolution transfer matrix $\mathbf{G}_t$, and $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t)$ the state evolution error. Both the scalar observational variance V_t, and the system evolution variance matrix $\mathbf{W}_t$, are assumed known (although in practice they rarely are - and the issues surrounding the ease of their specifications form most of chapters 3 and 4); additionally, note that we are free to express the evolution of the state vector entirely as we wish through equation 2.2.

Taking a state vector posterior at time $t-1$ (given all available information, D_{t-1}, up to time $t-1$) of $(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim \mathcal{N}(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$, for some known mean $\mathbf{m}_{t-1}$ and $p \times p$ variance matrix $\mathbf{C}_{t-1}$, we can then calculate a prior for the state vector at the next time point t from equation 2.2 - we know

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\theta}_t|D_{t-1}] &= \mathrm{E}[(\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t)|D_{t-1}] \\
&= \mathbf{G}_t\mathbf{m}_{t-1} \quad \text{since } \mathrm{E}[\mathbf{w}_t|D_{t-1}] = \mathbf{0} \ , \\
\text{and also } \mathrm{Var}(\boldsymbol{\theta}_t|D_{t-1}) &= \mathrm{Var}((\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t)|D_{t-1}) \\
&= \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{W}_t \ ,
\end{aligned}$$

since $\mathrm{Var}(\mathbf{w}_t|D_{t-1}) = \mathbf{W}_t$, and it is further assumed that the evolution error $\mathbf{w}_t$ is independent of the previous state vector. So, writing $(\boldsymbol{\theta}_t|D_{t-1}) \sim \mathcal{N}(\mathbf{a}_t, \mathbf{R}_t)$ for $\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{W}_t$, we then have the one step-ahead

forecast for Y_t, obtainable from the observation equation 2.1, of

$$\begin{aligned} \mathrm{E}[Y_t|D_{t-1}] &= \mathrm{E}[(\mathbf{F}_t^T\boldsymbol{\theta}_t + v_t)|D_{t-1}] \\ &= \mathbf{F}_t^T\mathbf{a}_t \quad \text{since } \mathrm{E}[v_t|D_{t-1}] = 0, \\ \text{and } \mathrm{Var}(Y_t|D_{t-1}) &= \mathrm{Var}((\mathbf{F}_t^T\boldsymbol{\theta}_t + v_t)|D_{t-1}) \\ &= \mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t \ , \end{aligned}$$

since $\mathrm{Var}(v_t|D_{t-1}) = V_t$, and this time it is assumed that the observational error v_t is independent of the current state vector. We write $(Y_t|D_{t-1}) \sim \mathcal{N}(f_t, Q_t)$, where $f_t = \mathbf{F}_t^T\mathbf{a}_t$ and $Q_t = \mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t$.

The prior and one step-ahead forecast are then both made use of, upon receiving the next data point Y_t, to update the state vector. All information up to time t is now available - i.e. we have $D_t = \{Y_t, D_{t-1}\}$ - and therefore we can calculate the updated posterior that constitutes the valuable Kalman Filter; this is defined via the posterior equations of

$$(\boldsymbol{\theta}_t|D_t) \sim \mathcal{N}(\mathbf{m}_t, \mathbf{C}_t), \tag{2.3}$$

$$\begin{aligned} \text{for } \mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t e_t \\ \mathbf{C}_t &= \mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t^T Q_t \ , \end{aligned}$$

$$\text{where further } \mathbf{A}_t = \frac{\mathbf{R}_t\mathbf{F}_t}{Q_t}, \text{ the adaptive coefficient vector at time } t,$$

$$\text{and } e_t = Y_t - f_t, \text{ the observed one step} - \text{ahead forecast error.}$$

This posterior is then used to provide the next prior $(\boldsymbol{\theta}_{t+1}|D_t)$, and so the

cycle repeats; at any stage we can subjectively interact with our prior to produce alterations in the one step-ahead forecast for Y_t, in the light of any relevant information that may have arisen. The only prerequisite of the system is that we 'kick-start' it by defining initial priors $\mathbf{m}_0$ and $\mathbf{C}_0$ such that $(\boldsymbol{\theta}_0|D_0) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0)$. These are chosen purely on the basis of the initial available information D_0, which may or may not include some data already, and will - almost by definition - usually include the subjective opinions of the practitioner on the nature of the data evolution. Very often it is not the choice of $\mathbf{m}_0$ that is the difficulty, but more the expression of uncertainty in this initial state, namely $\mathbf{C}_0$. This is undoubtedly a conceptually complicated area, and one which accordingly has many differing approaches, including the well-known option of choosing the rather paradoxically named 'uninformative prior'. This is not the place to open such a can of worms, and so it suffices to dwell merely upon two features of the initial prior specification. Firstly, and not entirely facetiously, we are actually choosing initial *posteriors* $(\mathbf{m}_0, \mathbf{C}_0)$, from which the first prior $(\boldsymbol{\theta}_1|D_0) \sim \mathcal{N}(\mathbf{G}_1\mathbf{m}_0, \mathbf{G}_1\mathbf{C}_0\mathbf{G}_1^T + \mathbf{W}_1)$ is then calculated; secondly, once we have established $\mathbf{m}_0$ and $\mathbf{C}_0$ we can run a speculative "what-if?" analysis of looking at the k^{th} step-ahead forecasts in the light of no data (other than D_0, which may or may not be the empty set).

The first point has been raised simply to emphasise the important conceptual difference between priors and posteriors within the DLM ('initial posteriors' was never likely to be used as the phrase is somewhat oxymoronic!), for in choosing $(\mathbf{m}_0, \mathbf{C}_0)$ it is vital to remember that we are *not* attempting to forecast the first (or next) data point in the time series under analysis, but instead are making a

statement about our current state. This is particularly relevant when initialising an analysis of seasonal data, where it is crucial to order the seasonal components correctly within the state vector $\mathbf{m}_0$, remembering that if our first data point is from, say, January (in monthly data), then the first seasonal component in $\mathbf{m}_0$ must correspond to December.

The second point, in relation to running a speculative "what-if?" analysis, is an important illustration of not only our new-found ability to operate either in the absence of any data whatsoever, or with merely a few previous observations, but also how we can check the suitability of the chosen model to our perceived data evolution. The k^{th} step-ahead forecast, from time t, is given by

$$f_{t+k,\,k} = \mathrm{E}[Y_{t+k}|D_t] = \mathbf{F}_{t+k}^T \mathbf{G}_{t+k} \mathbf{G}_{t+k-1} \ldots \mathbf{G}_{t+1} \mathbf{m}_t \ ,$$

and the evolution of this forecast function (calculating the associated variances of each forecast, $Q_{t+k,\,k}$, is equally simple) is evidently a potentially useful guide to the appropriateness of the model, in the light of prior knowledge of, or opinions on, the nature of the data evolution.

The posterior updating equation set 2.3 can be proven from either Bayes' theorem - this is done directly, via longhand substitution of the relevant Normal distributions into $p(\boldsymbol{\theta}_t|D_t) \propto p(\boldsymbol{\theta}_t|D_{t-1})p(Y_t|\boldsymbol{\theta}_t)$ - or from using standard bivariate Normal distribution theory. Either way, we prove a result which is the most vital innovation in the DLM framework; the source of its very dynamism, giving us the ability to update our prior beliefs and interact with the model (should the need arise) before and after receiving each new data point. And so

we come to the powerful Kalman Filter.

2.3 The Kalman Filter

The Kalman Filter (KF) was derived in 1960 in engineering journals by Kalman [22], and further in 1961 by Kalman and Bucy [23], to be deployed by various physical scientists in areas that were largely unknown territory to statisticians. Consequently its simplicity and value in time series analysis and recursive least squares algorithms (Young, 1974, [39]) was underappreciated and misunderstood, and this fear of the relatively unknown undoubtedly hindered the acceptance of the DLM following Harrison and Stevens' 1976 paper, which presented the KF as the simple statement of a black-box recursive estimation procedure. With hindsight, perhaps a more motivational derivation would have assisted with both clarity and the comprehension or interpretation of the many symbols within the KF updating procedure (Chatfield's criticism in the discussion of the Harrison and Stevens' paper was that there were "too many symbols"); the need for this simple motivation was emphasised by the appearance of a purely expository article on the KF a full seven years later by Meinhold and Singpurwalla [27], from which much of this section comes.

We start with our 'best guess' of $\boldsymbol{\theta}_t$ given information up to time $t-1$, namely the prior

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim (\mathbf{a}_t, \mathbf{R}_t) \ , \quad \mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1} \ ,$$

$$\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{W}_t \ .$$

Having made our one step-ahead forecast of Y_t, namely $f_t = \mathbf{F}_t^T \mathbf{a}_t$, and then having observed Y_t, we know the error in our prediction, $e_t = Y_t - f_t$. From Bayes' theorem, wanting the posterior $p(\boldsymbol{\theta}_t|D_t)$ is equivalent to requiring $p(Y_t|\boldsymbol{\theta}_t, D_{t-1})p(\boldsymbol{\theta}_t|D_{t-1})$, and since knowledge of Y_t is equivalent to knowledge of e_t, this can be written as

$$p(\boldsymbol{\theta}_t|D_t) \propto p(e_t|\boldsymbol{\theta}_t, D_{t-1})p(\boldsymbol{\theta}_t|D_{t-1}) \ , \tag{2.4}$$

where $(e_t|\boldsymbol{\theta}_t, D_{t-1}) \sim (\mathbf{F}_t^T(\boldsymbol{\theta}_t - \mathbf{G}_t\mathbf{m}_{t-1}), V_t)$ from the observation equation 2.1. So finally from Bayes' theorem,

$$p(\boldsymbol{\theta}_t|D_t) = \frac{p(e_t|\boldsymbol{\theta}_t, D_{t-1})p(\boldsymbol{\theta}_t|D_{t-1})}{\int_{(\text{all } \boldsymbol{\theta}_t)} p(e_t, \boldsymbol{\theta}_t|D_{t-1})\mathrm{d}\boldsymbol{\theta}_t} \ ,$$

which is, of course, exceptionally complicated to calculate!

It is possible to simplify the picture, though - firstly in the completely general case of an arbitrary prior distribution for Y_t, by using the appropriate conjugate prior analysis (see section 2.4), and secondly even further, by letting the prior distribution of $\boldsymbol{\theta}_t$ be Normal. In this case, taking the bivariate Normal distribution of

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right) \ , \tag{2.5}$$

we have that the conditional distribution of X_1 on X_2 is given by

$$(X_1|X_2 = x_2) \sim \mathcal{N}(\mu_1 + A_{12}A_{22}^{-1}(x_2 - \mu_2), A_{11} - A_{12}A_{22}^{-1}A_{21}) \ . \tag{2.6}$$

Conversely, the bivariate Normal distribution of X_1 and X_2 holds when $X_2 \sim$

$\mathcal{N}(\mu_2, A_{22})$ and the conditional distribution of X_1 on X_2 is as in 2.6. So, letting the prior $(\boldsymbol{\theta}_t|D_{t-1})$ take a Normal distribution, and identifying $X_1 = e_t$, $X_2 = \boldsymbol{\theta}_t$, $\mu_2 = \mathbf{G}_t\mathbf{m}_{t-1}$ and $A_{22} = \mathbf{R}_t$, we have from equation 2.6 that

$$(e_t|\boldsymbol{\theta}_t) \sim \mathcal{N}(\mu_1 + A_{12}\mathbf{R}_t^{-1}(\boldsymbol{\theta}_t - \mathbf{G}_t\mathbf{m}_{t-1}), A_{11} - A_{12}\mathbf{R}_t^{-1}A_{21}) \ ,$$

and equating this with the known $(e_t|\boldsymbol{\theta}_t, D_{t-1}) \sim \mathcal{N}(\mathbf{F}_t^T(\boldsymbol{\theta}_t - \mathbf{G}_t\mathbf{m}_{t-1}, V_t)$ gives $\mu_1 = 0$, $A_{12} = \mathbf{F}_t^T\mathbf{R}_t$, so that, further, symmetry ($A_{12} = A_{21}$) then gives $A_{11} = \mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t$. Hence our converse relationship, using equations 2.5 and 2.6, gives the bivariate Normal distribution of

$$\left(\begin{pmatrix} e_t \\ \boldsymbol{\theta}_t \end{pmatrix} \middle| D_{t-1} \right) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mathbf{G}_t\mathbf{m}_{t-1} \end{pmatrix}, \begin{pmatrix} \mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t & \mathbf{F}_t^T\mathbf{R}_t \\ \mathbf{R}_t\mathbf{F}_t & \mathbf{R}_t \end{pmatrix} \right) .$$

Thus, again from 2.5 and 2.6,

$$(\boldsymbol{\theta}_t|e_t, D_{t-1}) = (\boldsymbol{\theta}_t|D_t) \sim \mathcal{N}\Big(\mathbf{G}_t\mathbf{m}_{t-1} + \mathbf{R}_t\mathbf{F}_t(\mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t)^{-1}e_t, \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t(\mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t)^{-1} \Big) \ ,$$

and then we need only identify $\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$, $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_tQ_t^{-1}$ for $Q_t = \mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t$ to get the posterior given in the equation set 2.3.

It is now evident that what we have done, effectively, is construct the posterior mean as a regression function of $\boldsymbol{\theta}_t$ on the observed forecast error e_t, taking this mean to be the actual forecast plus a weighted proportion of the observed error. Additionally, this weighting term of $\mathbf{R}_t\mathbf{F}_t(\mathbf{F}_t^T\mathbf{R}_t\mathbf{F}_t + V_t)^{-1} = \mathbf{A}_t$ is the coefficient of least squares regression of $\boldsymbol{\theta}_t$ on e_t - in other words, our posterior

mean is equal to the one step-ahead forecast with a correction term added in precise relation to the performance of the forecast. This least squares estimation was the original derivation of the KF; the above Bayesian interpretation adds meaning and motivation to the updating procedure within the DLM.

Our derivation was solely possible due to the bivariate Normal distribution assumed for $\boldsymbol{\theta}_t$ and $e_t = Y_t$. The Bayesian updating becomes far more involved in non-Normal priors - indeed, to update the posterior via equation 2.4 above requires some impressive 'trickery' that we describe next.

2.4 Exponential families and conjugate priors

Many data series arise from observations that are evidently non-Normal. For continuous asymmetric data we can often work with transformations to produce symmetric distributions which are usefully modelled via Normality. However, where this is infeasible, especially in the case of discrete data distributions (as in the example that we follow later in the chapter), we must turn instead to the class of processes making use of the exponential family.

We can express most parent distributions for the data series Y_t as members of the general exponential family density

$$p(Y_t|\eta_t, V_t) = b(Y_t, V_t)\exp\{\frac{1}{V_t}(Y_t\eta_t - a(\eta_t))\} \ , \tag{2.7}$$

for some known functions $a(\eta_t)$ and $b(Y_t, V_t)$, the latter being a normalising constant, where V_t is the scale parameter (leading to the precision parameter $\phi_t = 1/V_t$) of the distribution, and where η_t is the natural parameter of the

distribution. The meanings of these quantities become more readily understood through studying examples; before we do so, note that

$$\begin{aligned} \mu_t = \mathrm{E}[Y_t|\eta_t, V_t] &= \frac{\mathrm{d}a(\eta_t)}{\mathrm{d}\eta_t} = \dot{a}(\eta_t) \\ \text{and } \mathrm{Var}(Y_t|\eta_t, V_t) &= V_t \frac{\mathrm{d}^2 a(\eta_t)}{\mathrm{d}\eta_t^2} = V_t \ddot{a}(\eta_t) \ , \end{aligned}$$

which is the motivation behind referring to $\phi_t = 1/V_t$ as the precision parameter - it is the scaling factor requires to give the variance of the process after calculation of $\ddot{a}(\eta_t)$.

Additionally, notice that we update to the posterior for η_t via Bayes' theorem, with $p(\eta_t|D_t) \propto p(\eta_t|D_{t-1})p(Y_t|\eta_t)$ (dropping the conditioning upon V_t, since we assume this to be known). With Normality of Y_t and the prior $p(\eta_t|D_{t-1})$, this calculation was simple; more generally it becomes manageable when we let the prior for η_t belong to a *conjugate family*. This prior density is of the form

$$p(\eta_t|D_{t-1}) = c(r_t, s_t)\exp\{s_t(x_t\eta_t - a(\eta_t))\} \ , \tag{2.8}$$

for $x_t = r_t/s_t$ the location, and s_t the precision parameter, of the prior, with $c(r_t, s_t)$ the known normalising constant (note the analogous definitions of s_t, x_t and $c(r_t, s_t)$ with the exponential family 2.7 earlier: $x_t \equiv Y_t$ and $s_t \equiv \phi_t$, with $c(r_t, s_t) \equiv b(Y_t, V_t)$). This conjugate prior allows us to calculate both the one step-ahead forecast distribution from $\int p(Y_t|\eta_t)p(\eta_t|D_{t-1})\mathrm{d}\eta_t$, namely

$$p(Y_t|D_{t-1}) = \frac{c(r_t, s_t)b(Y_t, V_t)}{c(r_t + \phi_t Y_t, s_t + \phi_t)} \ ,$$

and, most conveniently indeed, the updated posterior $p(\eta_t|D_t)$ - we have

$$\begin{aligned} p(\eta_t|D_t) &\propto p(\eta_t|D_{t-1})p(Y_t|\eta_t) \\ &\propto \exp\{r_t\eta_t - s_t a(\eta_t)\}\exp\{\phi_t(Y_t\eta_t - a(\eta_t))\} \\ &= \exp\{(r_t + \phi_t Y_t)\eta_t - (s_t + \phi_t)a(\eta_t)\} \ , \end{aligned}$$

which is of the same form as 2.8 above, with r_t updated to $r_t + \phi_t Y_t$, and s_t updated to $s_t + \phi_t$, yielding a new Normalising constant $c(r_t + \phi_t Y_t, s_t + \phi_t)$ which then gives the fully defined posterior

$$p(\eta_t|D_t) = c(r_t + \phi_t Y_t, s_t + \phi_t)\exp\{(r_t + \phi_t Y_t)\eta_t - (s_t + \phi_t)a(\eta_t)\} \ .$$

This completes the updating for the natural parameter of the model, but we generally wish to update fully for the posterior of $\boldsymbol{\theta}_t$ instead. So, note that defining $g(\eta_t) = \lambda_t = \mathbf{F}_t^T\boldsymbol{\theta}_t$, for some known function $g(.)$, gives

$$\begin{aligned} \mathrm{E}[\boldsymbol{\theta}_t|D_t] &= \mathrm{E}[\mathrm{E}[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}]|D_t] \\ \text{and } \mathrm{Var}(\boldsymbol{\theta}_t|D_t) &= \mathrm{Var}(\mathrm{E}[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}]|D_t) + \mathrm{E}[\mathrm{Var}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t] \ , \end{aligned}$$

but that we cannot calculate the conditional moments $\mathrm{E}[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}]$ and $\mathrm{Var}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})$ from the joint prior distribution of $\left(\begin{pmatrix} \lambda_t \\ \boldsymbol{\theta}_t \end{pmatrix}\middle| D_{t-1}\right)$, since we are no longer assuming a bivariate Normal distribution for this prior. Instead, we use a linear Bayesian estimation procedure which gives the optimal estimates of these con-

ditional moments as

$$\begin{aligned} \widehat{\mathrm{E}}[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}] &= \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t(\lambda_t - f_t)/q_t \\ \text{and } \widehat{\mathrm{Var}}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1}) &= \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t/q_t \ . \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{E}[\boldsymbol{\theta}_t|D_t] &= \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t(\mathrm{E}[\lambda_t|D_t] - f_t)/q_t = \mathbf{m}_t \\ \text{and } \mathrm{Var}(\boldsymbol{\theta}_t|D_t) &= \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t}{q_t^2}(\mathrm{Var}(\lambda_t|D_t)) + \mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t}{q_t} \\ \Rightarrow \mathbf{C}_t &= \mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t}{q_t}\left(1 - \frac{\mathrm{Var}(\lambda_t|D_t)}{q_t}\right) , \end{aligned}$$

and we can fully specify both $\mathrm{E}[\lambda_t|D_t] = \mathrm{E}[g(\eta_t)|D_t]$ and $\mathrm{Var}(\lambda_t|D_t) = \mathrm{Var}(g(\eta_t)|D_t)$ from the updated conjugate posterior $p(\eta_t|D_t)$. Thus we have a fully specified posterior $(\boldsymbol{\theta}_t|D_t) \sim (\mathbf{m}_t, \mathbf{C}_t)$ as desired.

Example. The example by which we choose to illustrate the conjugate prior analysis is the binomial case. Take Y_t to be binomial, i.e.

$$p(Y_t|\mu_t, n_t) = \begin{cases} \binom{n_t}{Y_t} \mu_t^{Y_t}(1-\mu_t)^{n_t - Y_t} & \text{for } Y_t = 0, 1, \ldots, n_t \\ 0 & \text{otherwise} , \end{cases}$$

where we have probability parameter μ_t such that $0 \leq \mu_t \leq 1$, and n_t trials such that $n_t > 0$. Then this is of the exponential family form 2.7 above, with $b(Y_t, V_t) = \binom{n_t}{Y_t}$, $y_t(Y_t) = Y_t/n_t$, $\eta_t = \ln(\mu_t/(1-\mu_t))$, $\phi_t = 1/V_t = n_t$, and $a(\eta_t) = \ln(1 + \mathrm{e}^{\eta_t})$. The relation for the natural parameter of the process, $\eta_t = \ln(\mu_t/(1-\mu_t))$, is known as the *logistic transform* of μ_t; notice that since

$0 \leq \mu_t \leq 1$ this transformation maps η_t to the whole of the real line. The conjugate prior for the probability parameter μ_t is then a Beta distribution, $(\mu_t|D_{t-1}) \sim \text{Beta}(r_t, s_t)$, namely

$$p(\mu_t|D_{t-1}) = \frac{\Gamma(r_t + s_t)}{\Gamma(r_t)\Gamma(s_t)} \mu_t^{r_t-1}(1-\mu_t)^{s_t-1} .$$

Using the logistic transform of μ_t for η_t then provides the form for the conjugate prior family of 2.8, but we can remain in the much simpler μ_t scale to update to a conjugate Beta posterior of

$$\begin{aligned} p(\mu_t|D_t) &\propto p(\mu_t|D_{t-1})p(Y_t|\mu_t, n_t) \\ &\propto \mu_t^{(r_t+Y_t)-1}(1-\mu_t)^{(s_t+n_t-Y_t)-1} \\ \Rightarrow p(\mu_t|D_t) &\sim \text{Beta}(r_t + Y_t, s_t + n_t - Y_t) . \end{aligned}$$

This has direct applications in any opinion poll analysis, for example, where the discrete data of positive respondents Y_t form a binomial sample, and we are interested in the underlying true proportion μ_t of these respondents, such as in the advertising awareness example that we look at now.

2.5 Example of the DLM

This example is taken directly from West and Harrison [36], chapter 14, and concerns the case study that appears there on advertising awareness of a particular product. We choose this case study to illustrate the non-Normal DLM and use of the conjugate prior analysis for the simple reason that it is the basis for

illustrations of subsequent discussions in chapter 3; it is also particularly useful for highlighting the framework and all-round application of the DLM. The full data set used with respect to this model in chapter 3 appears as Series 2 in the Appendix.

Briefly, a population survey is taken every week, where people are asked a standard question in relation to the advertising of a certain product (in this case a chocolate bar), and the number of positive respondents, Y_t, is recorded. During the preceding week the product will have been advertised on TV to a varying degree, and this level of advertising is calculated in standardised units known as TVR units, X_t (see Broadbent [5]). The DLM is defined here with the observation equation of

$$g(\eta_t) = \mu_t = \mathbf{F}_t^T \boldsymbol{\theta}_t \ ,$$

where $g(\eta_t)$ is the link function (a transformation 'linking' η_t to the real line), and is hence equal here to the inverse logistic transform due to the evidently binomial nature of the data series Y_t. Further, μ_t is interpreted as the underlying population response level for the data series Y_t - with $Y_t = n_t \mu_t$ for a population survey in week t of size n_t - and is the scale we work on. The state vector $\boldsymbol{\theta}_t$ is taken to be a 5-vector, so that $\boldsymbol{\theta}_t = (\alpha_t, \beta_t, \rho_t, \kappa_t, E_t)^T$, where the parameters represent the lower and upper thresholds of awareness, the memory decay rate, the penetration (of the advertising) parameter, and the effect of previous advertising at time t, respectively. Thus $\mathbf{F}_t^T = (1, 0, 0, 0, 1)$ in the observation equation - i.e. the mean response level μ_t is taken as the lower threshold plus

the current effect of previous advertising. Further, the evolution of the state is taken to be non-linear; each parameter is assumed to remain constant (up to the addition of the evolutionary noise), except for E_t, which evolves according to

$$E_t = (\beta_t - \alpha_t) - (\beta_t - \alpha_t - \rho_t E_{t-1})\mathrm{e}^{-\kappa_t X_t} \ .$$

Thus for no advertising in week t ($X_t = 0$), we have $E_t = \rho_t E_{t-1}$ (exponential decay in the effect of past advertising), whilst we obtain a fraction of the remaining awareness effect for each increase in X_t, ultimately requiring infinite advertising levels to obtain $E_t = \beta_t - \alpha_t$ (so that, in turn, $\mu_t = \beta_t$, the upper threshold level). This model definition, therefore, allows us to 'over-advertise' - we reach a point where we get little return in E_t for a large increase in X_t.

So, overall, we take the state equation of

$$\begin{aligned} \boldsymbol{\theta}_t &= g_t(\boldsymbol{\theta}_{t-1}) + \mathbf{w}_t \ , \\ \text{where } g(\mathbf{z}) &= (z_1, z_2, z_3, z_4, (z_2 - z_1) - (z_2 - z_1 - z_3 z_5)\mathrm{e}^{-z_4 X_t})^T \end{aligned}$$

for any 5-vector $\mathbf{z}$, and $\mathbf{w}_t \sim (\mathbf{0}, \mathbf{W}_t)$. This non-linear state evolution equation is all-well-and-good until we endeavour to evaluate the prior distribution $(\boldsymbol{\theta}_t | D_{t-1})$ - at this point, the state equation must be linearised as a first-order Taylor expansion of

$$\boldsymbol{\theta}_t \simeq g_t(\mathbf{m}_{t-1}) - \mathbf{G}_t \mathbf{m}_{t-1} + \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \mathbf{w}_t \ ,$$

for the $p \times p$ (where $\boldsymbol{\theta}_t$ is a p-vector) matrix $\mathbf{G}_t$ equal to

$$\mathbf{G}_t = \left[\frac{\delta g_t(\boldsymbol{\theta}_{t-1})}{\delta \boldsymbol{\theta}_{t-1}^T}\right]\Bigg|_{\boldsymbol{\theta}_{t-1}=\mathbf{m}_{t-1}} ,$$

namely

$$\mathbf{G}_t = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ e^{-\kappa_t X_t} - 1 & 1 - e^{-\kappa_t X_t} & E_{t-1} e^{-\kappa_t X_t} & X_t(\beta_t - \alpha_t - \rho_t E_{t-1}) e^{-\kappa_t X_t} & \rho_t e^{-\kappa_t X_t} \end{pmatrix}$$

in this example. Thus from a posterior $(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim (\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$ we can evaluate the prior

$$\begin{aligned} (\boldsymbol{\theta}_t|D_{t-1}) &\sim (\mathbf{a}_t, \mathbf{R}_t) , \\ \text{for } \mathbf{a}_t &= g_t(\mathbf{m}_{t-1}) \\ \text{and } \mathbf{R}_t &= \mathbf{G}_t \mathbf{C}_{t-1} \mathbf{G}_t^T + \mathbf{W}_t . \end{aligned}$$

The updating is in two stages, as usual - firstly, updating for μ_t is via the conjugate prior analysis outlined in section 2.4; taking the Beta prior for μ_t of $(\mu_t|D_{t-1}) \sim \text{Beta}(r_t, s_t)$, we have that $\text{E}[\mu_t|D_{t-1}] = \frac{r_t}{r_t+s_t}$. But further, $\text{E}[\mu_t|D_{t-1}] = f_t = \mathbf{F}_t^T \mathbf{a}_t$ which is fully known from the posterior of $\boldsymbol{\theta}_{t-1}$, as is $\text{Var}(\mu_t|D_{t-1}) = q_t = \mathbf{F}_t^T \mathbf{R}_t \mathbf{F}_t$, in turn equal to $\frac{f_t(1-f_t)}{r_t+s_t+1}$ from the Beta distribution for $(\mu_t|D_{t-1})$. Thus we can solve these two equations for r_t and s_t, yielding

$$r_t = f_t \left(\frac{f_t}{q_t}(1 - f_t) - 1\right)$$

$$\text{and} \quad s_t = (1-f_t)\left(\frac{f_t}{q_t}(1-f_t)-1\right) .$$

Making use of the conjugate prior analysis, we now know that the posterior $(\mu_t|D_t)$ is updated to a Beta$[r_t+Y_t, s_t+n_t-Y_t]$ distribution. So we can readily calculate

$$g_t = \mathrm{E}[\mu_t|D_t] = \frac{r_t+Y_t}{r_t+s_t+n_t}$$

$$\text{and} \quad p_t = \mathrm{Var}(\mu_t|D_t) = \frac{g_t(1-g_t)}{r_t+s_t+n_t+1} .$$

The second stage of the updating is then for $\boldsymbol{\theta}_t$; noticing that in this example, $\lambda_t = \mu_t = \mathbf{F}_t^T\boldsymbol{\theta}_t$, we can simply make use of the method of section 2.4 and evaluate both

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t(\mathrm{E}[\mu_t|D_t]-f_t)/q_t = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t(g_t-f_t)/q_t$$

$$\text{and} \quad \mathbf{C}_t = \mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t}{q_t}\left(1-\frac{\mathrm{Var}(\mu_t|D_t)}{q_t}\right) = \mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t^T\mathbf{R}_t}{q_t}\left(1-\frac{p_t}{q_t}\right) .$$

2.6 Discounting

So far in this chapter, we have introduced many concepts around the basic structure of the DLM, in order to cope with both non-Normal and non-linear models. However, one concept which is crucial to each and every one of these models arises in the sequential updating of the posterior $(\boldsymbol{\theta}_{t-1}|D_{t-1})$ to form prior values for $(\boldsymbol{\theta}_t|D_{t-1})$.

To recap, in general the DLM is defined via two main equations; the observation equation 2.1 given by

$$Y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + v_t ,$$

for $\mathbf{F}_t$ known and $v_t \sim \mathrm{N}(0, V_t)$, and the state or system equation 2.2

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \mathbf{w}_t ,$$

for $\mathbf{w}_t \sim \mathrm{N}(\mathbf{0}, \mathbf{W}_t)$, and uncorrelated with $\boldsymbol{\theta}_{t-1}$, which gives a recurrence for the updating of the state vector $\boldsymbol{\theta}_t$.

This then leads to our one-step ahead forecast for $\boldsymbol{\theta}_t$; for if we start with a posterior at time $t-1$,

$$(\boldsymbol{\theta}_{t-1} | D_{t-1}) \sim (\mathbf{m}_{t-1}, \mathbf{C}_{t-1}) ,$$

we now have that the prior for $\boldsymbol{\theta}_t$ is given, from the state equation and posterior, as

$$\begin{aligned} (\boldsymbol{\theta}_t | D_{t-1}) &\sim (\mathbf{a}_t, \mathbf{R}_t) , \\ \text{for} \quad \mathbf{a}_t &= \mathbf{G}_t \mathbf{m}_{t-1} \\ \text{and} \quad \mathbf{R}_t &= \mathbf{G}_t \mathbf{C}_{t-1} \mathbf{G}_t^T + \mathbf{W}_t . \end{aligned} \tag{2.9}$$

However, Ameen and Harrison [2] state that one of the major obstacles in the widespread adoption of Bayesian dynamic modelling - an issue raised

earlier at the beginning of section 2.2 - has been the difficulty of specifying this (not necessarily constant) system variance matrix $\mathbf{W}_t$. Even experienced practitioners have little feel for this matrix, as well as for the observational variance V_t. In addition, problems arise due to the non-uniqueness of $\mathbf{W}_t$, along with its invariance to the scale on which the independent variables are measured.

Their solution is to introduce the concept of discounting. The approach is to define a discount matrix $\mathbf{B}_t$ such that

$$\begin{aligned} \mathbf{R}_t &= \mathbf{B}_t^{-1/2}\mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T\mathbf{B}_t^{-1/2} , \\ \text{where } \mathbf{G}_t &= \text{diag}(\mathbf{G}_1, ..., \mathbf{G}_r) \\ \text{and } \mathbf{B}_t &= \{\delta_1\mathbf{I}_{n_1}, ..., \delta_r\mathbf{I}_{n_r}\} , \end{aligned} \tag{2.10}$$

each block $\mathbf{G}_i$ is of full rank n_i, $0 < \delta_i \leq 1$ for all $i = 1, ..., r$, and $\mathbf{I}_{n_i}$ is the identity matrix of dimension n_i.

This alternative to specifying the system variance matrix $\mathbf{W}_t$ possesses two desirable properties of forecasting - ease of model application, and conceptual parsimony. In addition, Ameen and Harrison state that many forecasters have a 'natural feel' for the set of discounting factors $\{\delta_1, ..., \delta_r\}$. It is also subsequently possible to apply established methods for estimation of the observation variance, V_t, once the discount factors are chosen.

However, upon further examination, attempting to specify either $\mathbf{W}_t$ *or* a discounting matrix $\mathbf{B}_t$ can be seen to have certain flaws which lead to practitional difficulties, making the decision of whether to employ a discounting

matrix or not very much problem specific, and dependent on the defining equations of the DLM in question. We will develop further theoretical limiting results for the DLM in the rest of this chapter, enabling us to understand where both approaches have shortcomings - the subsequent discussion in chapter 3.

2.7 Time series DLMs

In many DLMs the two defining equations, 2.1 and 2.2, will be concerned with constant $\mathbf{F}_t = \mathbf{F}$ and $\mathbf{G}_t = \mathbf{G}$ (all polynomial trend models, for instance; see chapter 3 for examples). These model specifications are known as Time Series DLMs (TSDLM), defined in shorthand by the quadruple $\{\mathbf{F}, \mathbf{G}, V_t, \mathbf{W}_t\}$. Further, all classical stationary time series can be expressed as *constant* TSDLMs, defined by $\{\mathbf{F}, \mathbf{G}, V, \mathbf{W}\}$, as we shall see shortly. Although we have already discussed the restrictiveness of this class of models under the *Box-Jenkins* method of analysis, representing them as constant TSDLMs allows us to combat nearly all of the four main restrictions mentioned in section 2.1. We are no longer isolated from the model, being instead able to intervene at any stage of the analysis and input information pertaining to the data evolution; additionally we are able to cope with little or no data, since by simply specifying initial priors $(\mathbf{m}_0, \mathbf{C}_0)$ (which may or may not be based upon previous information) it is possible to run speculative 'what-if?' analyses in the absence of any future data. And by looking at these simpler models practitioners gain insight to more complicated forecasting systems, obtained by the superposition of two-or-more such TSDLMs.

2.8 Observability

As with all statistical analyses, it is desirable to ensure that we specify a parsimonious model. With respect to a TSDLM, overparameterisation can be avoided by checking that the defined model is observable. Although in what follows we continue to adopt West and Harrison's notation and approach, the concept of observability is attributable to Kalman (see, for instance, Kalman et al. [24]).

Defining the mean response function, $\mu_{t+k} = \mathrm{E}[Y_{t+k}|\boldsymbol{\theta}_{t+k}] = \mathbf{F}^T\boldsymbol{\theta}_{t+k}$, and the forecast function $f_t(k) = \mathrm{E}[\mu_{t+k}|D_t] = \mathbf{F}^T\mathbf{G}^k\mathbf{m}_t$, we can then define the p-vector $\boldsymbol{\mu}_t = (\mu_t, \mu_{t+1}, \ldots, \mu_{t+p-1})^T$ (remembering that $\boldsymbol{\theta}_t$ is p-dimensional), so that $\boldsymbol{\mu}_t = \mathbf{T}\boldsymbol{\theta}_t$, where

$$\mathbf{T} = \begin{pmatrix} \mathbf{F}^T \\ \mathbf{F}^T\mathbf{G} \\ \vdots \\ \mathbf{F}^T\mathbf{G}^{p-1} \end{pmatrix}$$

is a $p \times p$ matrix, known as the observability matrix. We now require that μ_t should contain sufficient information to provide exact knowledge of the state $\boldsymbol{\theta}_t$ (given that we also know the state evolution errors $\mathbf{w}_{t+i}$, $i = 0, \ldots, p-1$); i.e. it should be possible to calculate $\boldsymbol{\theta}_t = \mathbf{T}^{-1}\boldsymbol{\mu}_t$. Hence *Kalman's observability criterion* is that $\mathbf{T}$ must be non-singular, and we say that under this condition the TSDLM $\{\mathbf{F}, \mathbf{G}, V, \mathbf{W}\}$ is *observable.*

It is worth noting that any p-dimensional model, for $p > 1$, in which $\mathbf{G}$ is the identity matrix, is evidently not observable - the observability matrix becomes

$\mathbf{T} = \begin{pmatrix} \mathbf{F}^T \\ \vdots \\ \mathbf{F}^T \end{pmatrix}$ and of rank 1. In general, we can reparametrise any unobservable model in which $\mathbf{T}$ is of rank, say, $p - r$, via a linear transformation to an observable model of dimension exactly $p - r$.

However, it is also notable that any standard seasonality model is also unobservable; taking the model

$$\{(1, \mathbf{E}_{p-1})^T, \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{P} \end{pmatrix}, V, \mathbf{W}\},$$

where $\mathbf{E}_{p-1} = (1, 0, \ldots, 0)^T$ is $(p-1)$-dimensional, and

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ 0 & 0 & \ldots & \ldots & \ldots & 1 \\ 1 & 0 & \ldots & \ldots & \ldots & 0 \end{pmatrix}$$

is the $(p-1) \times (p-1)$ permutation matrix - so that the data series $\{Y_t\}$ is seen as an underlying mean level plus a seasonality component, with the seasonality having period $p - 1$ time points - results in $\mathbf{P}^{p-1} = \mathbf{I}_{p-1}$, the $(p-1) \times (p-1)$ identity matrix. This therefore results in $\mathbf{G}^{p-1} = \mathbf{I}_p$, from whence $\mathbf{F}^T\mathbf{G}^{p-1} = \mathbf{F}^T$ and so

$$\mathbf{T} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & \ldots & 1 \\ 1 & 1 & 0 & \ldots & \ldots & 0 \end{pmatrix}.$$

Hence $\mathbf{T}$ has rank $p - 1$ and this seasonality model can be reparametrised to

a $(p-1)$-dimensional model (by incorporating the underlying mean level into each seasonality term, in effect). What most practitioners will wish to do in this DLM, however, is to constrain the seasonality components to having zero total sum; i.e. if $\boldsymbol{\theta}_t = (\mu_t, \boldsymbol{\phi}_t)^T$, for μ_t the mean level and $\boldsymbol{\phi}_t$ a $(p-1)$-vector of seasonality components, we would want $\mathbf{1}^T\boldsymbol{\phi}_t = 0$ for all t, or equivalently $(0,1,1,\ldots,1)\boldsymbol{\theta}_t = 0$. This leads to the idea of *constrained* observability - taking the unobservable model $\{\mathbf{F}, \mathbf{G}, V, \mathbf{W}\}$, where further there is the constraint $\mathbf{C}\boldsymbol{\theta}_t = \mathbf{c}$ for known $\mathbf{C}$ and $\mathbf{c}$, the TSDLM is called constrained observable iff the extended observability matrix $\mathbf{T} = \begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}$ is of full rank p.

It is now possible to show (West and Harrison [36], pps. 148-150) that any such observable constant TSDLM, with V and $\mathbf{W}$ both finite, has the limiting form of

$$\lim_{t\to\infty}\{\mathbf{A}_t, \mathbf{C}_t, \mathbf{R}_t, Q_t\} = \{\mathbf{A}, \mathbf{C}, \mathbf{R}, Q\} \ .$$

The proof of this special case of stationarity is quite complex; in general it is impractical to attempt to express the form for the limiting value of $\mathbf{C}_t \xrightarrow[\infty]{t} \mathbf{C}$ algebraically in terms of the variances V and $\mathbf{W}$. Instead, it is worthy of note that from equation set 2.3 the evolution of the uncertainty in the state, $\mathbf{C}_t$, is independent of the data series $\{Y_t\}$ in all linear DLMs. Here, in the constant TSDLM, the convergent value of $\mathbf{C}$ is therefore predetermined solely by the specification of V and $\mathbf{W}$ (and is independent of the initial prior $\mathbf{C}_0$), and so rather than endeavouring to solve for $\mathbf{C}$ from

$$\mathbf{C} = \mathbf{R} - \mathbf{A}\mathbf{A}^T Q = \mathbf{G}\mathbf{C}\mathbf{G}^T + \mathbf{W} - \frac{(\mathbf{G}\mathbf{C}\mathbf{G}^T + \mathbf{W})\mathbf{F}\mathbf{F}^T(\mathbf{G}\mathbf{C}\mathbf{G}^T + \mathbf{W})^T}{\mathbf{F}^T(\mathbf{G}\mathbf{C}\mathbf{G}^T + \mathbf{W})\mathbf{F} + V} \ ,$$

we can simply obtain a numerical value for $\mathbf{C}$ from iteration of $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t^T Q_t$.

Having found $\mathbf{C}$, the other limiting values of $\mathbf{R}, \mathbf{A}$ and Q are obtainable from the defining equations of the Kalman Filter (2.3). Then, finally, from this limiting form it is always possible to express the constant TSDLM in a general model form of

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + e_t + \sum_{j=1}^{p} \beta_j e_{t-j} \ , \tag{2.11}$$

where the α_i's are determined from functions of the eigenvalues of $\mathbf{G}$, and the $-\beta_i$'s from the same functions of the eigenvalues of $\mathbf{H} = (\mathbf{I} - \mathbf{A}\mathbf{F}^T)\mathbf{G} = \mathbf{C}\mathbf{R}^{-1}\mathbf{G}$. If, further, the e_t are $\overset{iid}{\sim}\mathcal{N}(0, \sigma_e^2)$, then this model representation is evidently an ARIMA$(0, p, p)$ process; however, we do not have to make this general assumption about the distribution of the one step-ahead forecast errors (and it will not hold initially in any model) to derive this result. Therefore it is true to say that through this representation we can describe all general ARIMA processes as particular constant TSDLMs, but that the converse is *not* true, a point stressed by Harrison and Akram [17]. This more general model form 2.11 will be most useful at the start of chapter 4, when we endeavour to solve for $\mathbf{W}$.

2.9 Canonical equivalence

Having seen that via a particular linear transformation any unobservable TSDLM can be reparamaterised as a simpler, observable model of lesser dimension, and having seen the convergence properties of such TSDLMs, we will usually

only deal with these (sometimes constrained) observable models. However, under the broad heading of observability, there are many models which have similar properties, and it is useful to group together these TSDLMs into classes linked by their similarities.

The desired property with which we work is the forecast function $f_t(k)$ of section 2.2. We say that two models M and M_1, defined by the quadruples $\{\mathbf{F}, \mathbf{G}, V_t, \mathbf{W}_t\}$ and $\{\mathbf{F}_1, \mathbf{G}_1, V_{t1}, \mathbf{W}_{t1}\}$, with observability matrices $\mathbf{T}$ and $\mathbf{T}_1$ respectively, are *similar* iff $\mathbf{G}$ and $\mathbf{G}_1$ have identical eigenvalues. This then leads to identification of a matrix $\mathbf{H}$ - and it is readily shown that $\mathbf{H} = \mathbf{T}^{-1}\mathbf{T}_1$ - such that $\mathbf{G} = \mathbf{H}\mathbf{G}_1\mathbf{H}^{-1}$ and $\mathbf{F}^T = \mathbf{F}_1^T\mathbf{H}^{-1}$, and hence to the reparametrisation $\boldsymbol{\theta}_{t1} = \mathbf{H}^{-1}\boldsymbol{\theta}_t$, whence

$$\begin{aligned} Y_t &= \mathbf{F}_1^T\mathbf{H}^{-1}\boldsymbol{\theta}_t + v_{t1} \\ \text{and } \boldsymbol{\theta}_t &= \mathbf{H}\mathbf{G}_1\mathbf{H}^{-1}\boldsymbol{\theta}_{t-1} + \mathbf{H}\mathbf{w}_{t1} \end{aligned}$$

from substitution for $\boldsymbol{\theta}_{t1}$ in equations 2.1 and 2.2. Therefore

$$\begin{aligned} f_{t1}(k) &= \mathbf{F}_1^T\mathbf{G}_1^k\mathbf{m}_{t1} = (\mathbf{F}_1^T\mathbf{H}^{-1})(\mathbf{H}\mathbf{G}_1^k\mathbf{H}^{-1})(\mathbf{H}\mathbf{m}_{t1}) \\ &= \mathbf{F}^T\mathbf{G}^k\mathbf{H}\mathbf{m}_{t1}\ , \end{aligned}$$

and $f_{t1}(k)$ will be equal to $f_t(k)$ - i.e. the two models M and M_1 will have identical forecast distributions - iff $V_{t1} = V_t$, $\mathbf{H}\mathbf{W}_t\mathbf{H}^T = \mathbf{W}_{t1}$, with further $\mathbf{m}_t = \mathbf{H}\mathbf{m}_{t1}$ and $\mathbf{C}_t = \mathbf{H}\mathbf{C}_{t1}\mathbf{H}^T$. This is our definition of equivalence (written $M \equiv M_1$).

If we next define the $p \times p$ Jordan block

$$\mathbf{J}_p(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda & 1 & \dots & \dots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \dots & 1 \\ 0 & 0 & 0 & \dots & \dots & \lambda \end{pmatrix},$$

it is easy to show that any observable TSDLM with system matrix $\mathbf{G}$, having a single eigenvalue λ of multiplicity p, is similar to the TSDLM with system matrix $\mathbf{J}_p(\lambda)$. Further, the first element of $\mathbf{F}$ must be non-zero for the TSDLM to be observable, and so the simplest representation of this particular TSDLM is the canonically similar model $\{\mathbf{E}_p, \mathbf{J}_p(\lambda), V_t, \mathbf{W}_t\}$, where $\mathbf{E}_p = (1, 0, \dots, 0)^T$ is a p-vector.

This result generalises to multiple real eigenvalues: any observable TSDLM $\{\mathbf{F}, \mathbf{G}, V_t, \mathbf{W}_t\}$ in which $\mathbf{G}$ has s distinct real eigenvalues $\lambda_1, \dots, \lambda_s$ each with multiplicities $r_1, \dots, r_s$ respectively, and which has observability matrix $\mathbf{T}$, is canonically similar to the model $\{\mathbf{E}, \mathbf{J}, V_t, \mathbf{W}_t\}$ where $\mathbf{E} = (\mathbf{E}_{r_1}, \dots, \mathbf{E}_{r_s})^T$ and $\mathbf{J} = \text{blockdiag}[\mathbf{J}_{r_1}(\lambda_1), \dots, \mathbf{J}_{r_s}(\lambda_s)]$, with observability matrix $\mathbf{T}_0$. If, additionally, we have $\mathbf{W}_{t1} = \mathbf{H}\mathbf{W}_t\mathbf{H}^T$ for $\mathbf{H} = \mathbf{T}_0^{-1}\mathbf{T}$, then the TSDLM as defined is *canonically equivalent* to $\{\mathbf{E}, \mathbf{J}, V_t, \mathbf{W}_{t1}\}$. This does, incidentally, generalise yet further to multiple complex eigenvalues of $\mathbf{G}$, but we shall concern ourselves with just the above case.

Example. In our earlier general seasonal model with underlying mean level, and taking the 3-dimensional model of

$$M = \left\{(1,1,0)^T, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, V_t, \mathbf{W}_t\right\},$$

we have that $\mathbf{G}$ has eigenvalues -1 and 1, with multiplicities 1 and 2. Thus its canonically similar form is

$$M_1 = \left\{(1,1,0)^T, \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, V_{t1}, \mathbf{W}_{t1}\right\},$$

with observability matrices

$$\mathbf{T} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$$\text{and } \mathbf{T}_0 = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Note that as model M stands, it is not observable; if we calculate the similarity matrix $\mathbf{H} = \mathbf{T}_0^{-1}\mathbf{T}$ between M and M_1 we find

$$\mathbf{H} = \begin{pmatrix} 0 & 1/2 & -1/2 \\ 1 & 1/2 & 1/2 \\ 0 & 0 & 0 \end{pmatrix},$$

which then suggests that the canonically equivalent form for M has $\mathbf{W}_{t1} = \mathbf{H}\mathbf{W}_t\mathbf{H}^T$ of the form $\begin{pmatrix} \mathbf{W}'_{t1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$, where $\mathbf{W}'_{t1}$ is a 2×2 matrix. This in turn suggests the possible reparametrisation to a 2-dimensional model, which we avoid by adding the constraint $(0,1,1)\boldsymbol{\theta}_t = 0$ (the zero sum seasonality constraint) so that M is now constrained observable, with observability matrix

$$\mathbf{T}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

$$\Rightarrow \mathbf{H}_1 = \mathbf{T}_0^{-1}\mathbf{T}_1 = \begin{pmatrix} 1/4 & 3/4 & 1/4 \\ 3/4 & -3/4 & -1/4 \\ -1/2 & 1/2 & 1/2 \end{pmatrix},$$

and thus the canonically equivalent form is

$$M_1 = \left\{(1,1,0)^T, \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, V_t,\ \mathbf{H}_1\mathbf{W}_t\mathbf{H}_1^T\right\},$$

for $\mathbf{H}_1$ defined as above.

2.10 Long-term equivalence

The canonical representation defined in section 2.9 is useful for reducing any given model to a simple-to-understand, standard form, and saves us replication of effort when analysing two models which appear different at first sight of their definitions, but in fact turn out to be canonically equivalent. However, in chapter 3 we shall be interested in models which display equivalence of a subtle variation on canonical equivalence. This is when two models , *not necessarily canonically equivalent*, still have identical limiting forms of forecast distribution.

Definition. Defining two models, written in canonically similar form as

$$M = \{\mathbf{E}, \mathbf{J}, V_t, \mathbf{W}_t\}$$

$$\text{and } M' = \{\mathbf{E}, \mathbf{J}, V_t, \mathbf{W}_t + \delta\mathbf{W}_t\} ,$$

we say that M and M' are *long-term equivalent* if they have identical convergent forecast distributions.

Note that M and M' are no longer necessarily canonically equivalent, as there is no guarantee that we shall be able to write the variance matrix $\mathbf{W}_t+\delta\mathbf{W}_t$ as $\mathbf{H}_1\mathbf{W}_t\mathbf{H}_1^T$, in terms of the similarity matrix $\mathbf{H}_1$. Indeed, it is highly probable that the two models M and M' will belong to distinct canonically equivalent classes. However, we shall see in chapter 3 that in any constant TSDLM, written in canonical form, it is always possible to solve partially for the perturbation matrix $\delta\mathbf{W}_t$, leaving this matrix with $\frac{1}{2}p(p-1)$ degrees of freedom. Hence, once we have reduced our model of interest to canonical form, we can find an infinite class of long-term equivalent models, not necessarily restricted by having the same canonically equivalent class, which is defined instead by a particular limiting forecast distribution. This has many fundamental implications which we explore in depth later in chapter 3, and which motivates all of chapter 4.

Having developed so much theory, we can now turn attention to the issues raised earlier in section 2.6, relating to the discounting debate.

Chapter 3

The Discounting Debate

3.1 Sensitivity of the DLM to choice of V_t and $\mathbf{W}_t$

When any inexperienced practitioners first attempt an analysis utilising a DLM, they are very likely to face the majority of their work - even though they may not realise it - combatting a seemingly simple task; one which is not in defining the model state $\boldsymbol{\theta}_t$ or its evolution matrix $\mathbf{G}_t$, nor the independent variables in $\mathbf{F}_t$. Instead, it lies in specifying the observational variance V_t and the system error variance $\mathbf{W}_t$. And it quickly becomes apparent to the practitioner that the resulting performance of the DLM is very sensitive indeed to the choice of these variances - specifically, in their choice relative to one another. If one chooses elements too large in $\mathbf{W}_t$ with respect to V_t, all variations in the data series are accounted for as underlying changes in the state vector, and hence the forecasts tend to follow the last data point. Conversely, specifying V_t too large

with respect to some or all of the elements in $\mathbf{W}_t$ results in data perturbations being seen as merely natural noise in the time series, and so the forecasts tend to ignore the data altogether, following instead an evolution similar to a 'what-if?' analysis of considering the forecasts from initial priors alone. West and Harrison [36] provide a more detailed discussion of this, specifically in relation to the first-order polynomial model and sensitivity with respect to the signal-to-noise ratio $r = W/V$.

However, Ameen and Harrison's [2] introduction of the discounting matrix $\mathbf{B}_t$, in equation 2.10, goes a long way towards tackling this problem. The analysis is still slightly sensitive to the choices of the discounting factors δ_i, but to a much lesser degree; further, it is easy for a practitioner to realise, from following a few analyses that utilise a discounting matrix, that the δ_i are almost always chosen as being greater than 0.85-or-so, thus effectively self-selecting $\mathbf{B}_t$.

This 'ease of model application', as Ameen and Harrison put it, is very luring for the unwary. However, just because the practitioner has a 'natural feel' for the discount factors (i.e. make them around 0.95, give-or-take), he must not be tempted into its indiscriminant use. The advantages are not so clear.

3.2 Discussion of the discounting approach

There is undoubtedly a physical need to build in a 'system equation error', $\mathbf{w}_t$, in our parameter updating process. This need is easily interpretable and understandable, for it is simply a comment on the non-uniformity of the physical laws governing our model. When specifying $\mathbf{W}_t$ we are attempting to attach a

likely range of values to this system error.

Let us now turn attention to the interpretation of the discount matrix, $\mathbf{B}_t$. Ameen and Harrison state that "a single discount factor δ describes the rate at which information is lost with time so that, if the current information is now worth I units, then its worth with respect to a period k steps ahead is $\delta^k I$ units". It is then subsequently correctly noted in West and Harrison [36] that discounting should only be thought of as applicable for one-step ahead forecasting - i.e. for $k = 1$ - since the evolution of loss of information is evidently an additive one from equation 2.9, and not the exponential decay as suggested by Ameen and Harrison. Indeed, West and Harrison continue to state that when looking at 'what-if?' scenarios, where practitioners are interested in long-term forecasting from their present beliefs, we are faced with no choice but to revert to constructing a constant matrix $\mathbf{W}_t$ which we will sequentially add to our updating of $\mathbf{R}_t(k)$ for all $k = 1, 2, ...$ ($\mathbf{R}_t(k)$ is the k^{th} step-ahead value of $\mathbf{R}_t$). Their solution to our sudden problem is to construct $\mathbf{W}_t$ using our current discount matrix $\mathbf{B}_t$; for if δ_i is the discount factor associated with the i^{th} diagonal block in $\mathbf{G}_t$, then the corresponding i^{th} block component in $\mathbf{W}_t$ is given as

$$\mathbf{W}_{it} = \mathbf{G}_{it}\mathbf{C}_{i,t-1}\mathbf{G}_{it}^T(1 - \delta_i)/\delta_i \ , \quad (i = 1, ..., r).$$

But what price do we pay? If we are discounting the i^{th} block by an (arbitrary) $100(1-\delta_i)\%$ now, and we wish (as is often the case) to speculate ahead for a relatively long period of time (such as for an entire year whilst working with weekly data), we are faced with the distinct possibility of either meaninglessly

large variances and covariances in $\mathbf{R}_t(k)$ (arising from currently small amounts of knowledge and/or arbitrary selection of a relatively small δ_i), or extremely small values in $\mathbf{R}_t(k)$ representing unrealistically accurate beliefs (which would occur when current knowledge levels are high, so that $100(1-\delta_i)\%$ of $\mathbf{R}_{it}$ is extremely small and even 50 sequential additions of it increase $\mathbf{R}_t(k)$ values by only a small level). We are also faced with the same problems when dealing with missing data within a time series, where a possibly long break in data collecting could see either extreme occurring in our prior knowledge at the next available data point. The former of these scenarios is of particular importance where some or all of the parameters of the model are bounded in range (as in the later example, where all parameters must lie within (0,1) - inflating variances here soon become ridiculously large with respect to this measurement scale); the latter of the two scenarios is far more serious where future forecasting performance is concerned, for it will take the model a long time to readjust to any perturbations in its parameter values that may have occurred during the period of no data.

This brings us to the most worrying problem of the discounting approach. More and more information about our state vector is represented by smaller and smaller variances (and covariances alike) in $\mathbf{R}_t$, culminating in the very important special case of *full* knowledge of the state at time $t-1$, in which case $\mathbf{C}_{t-1} = \mathbf{0}$ and the discounting approach breaks down completely. (This is just as relevant even if we only have full knowledge of just one parameter θ_i, where $\mathbf{C}_{i,t-1}$ is 0). Despite it only being relevant specifically at the initialisation of a DLM (when we specify $\mathbf{m}_0$ and $\mathbf{C}_0$), or in any non-linear model, this is still not

a negligible problem, for it is most paradoxical that given a situation where we should be at our most powerful (i.e. with full knowledge about the present), we are reduced to a non-functioning level.

3.3 Example

This most worrying aspect of discounting is best illustrated by reference to the non-linear example described earlier in section 2.5. To compensate for a lack of experience in the field of advertising awareness modelling, the first 75 data points taken are exactly as those appearing in West and Harrison, as is the analysis over this time, which makes use of the discounting approach. Hence possible problems arising from naive initial priors and choice of discount parameters are avoided. The full data sets are given in Table 3.1.

Table 3.1: Advertising awareness data: X_t and $\mu_t = Y_t/n_t$.

X_t : TVR units (weekly, by row)														
0.05	0.00	0.20	7.80	6.10	5.15	1.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.50	4.60	3.70	1.45	1.20	2.00	3.40	4.40	3.80	3.90	5.00	0.10	0.60	3.85	3.50
3.15	3.30	0.35	0.00	2.80	2.90	3.40	2.20	0.50	0.00	0.00	0.10	0.85	4.65	5.10
5.50	2.30	4.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.0	10.0	9.00	8.00	7.00	6.00
5.00	4.00	3.00	2.00	1.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.0
10.0	9.00	8.00	7.00	6.00	5.00	4.00	4.00	5.00	6.00	7.00	8.00	9.00	10.0	11.0
12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	20.0	19.0	18.0	17.0	16.0	15.0
14.0	13.0	12.0	11.0	11.0	12.0	2.90	4.47	2.24	6.71	3.22	7.29	6.66	2.48	3.64
8.70	7.11	2.30	6.40	3.20	7.11	3.57	7.93	5.13	6.40	3.31	2.68	5.39	7.22	4.40
6.69	8.69	6.32	5.99	6.94	6.19	7.59	3.59	8.44	8.61	8.08	9.32	9.13	8.39	9.58

μ_t : Awareness response proportion, Y_t/n_t															
0.40	0.41	0.31	0.40	0.45	0.44	0.39	0.50	0.32	0.42	0.33	0.24	0.25	0.32	0.28
0.25	0.36	0.38	0.36	0.29	0.43	0.34	0.42	0.50	0.43	0.43	0.52	0.45	0.30	0.55
0.33	0.32	0.39	0.32	0.30	0.44	0.27	0.44	0.30	0.32	0.30	0.00	0.00	0.00	0.33
0.48	0.40	0.44	0.40	0.34	0.37	0.37	0.23	0.30	0.21	0.23	0.22	0.25	0.23	0.14
0.21	0.16	0.19	0.07	0.26	0.16	0.21	0.07	0.22	0.10	0.15	0.15	0.22	0.11	0.14
0.04	0.19	0.19	0.29	0.36	0.40	0.28	0.43	0.57	0.58	0.59	0.67	0.50	0.63	0.66
0.61	0.48	0.65	0.30	0.50	0.41	0.51	0.36	0.44	0.47	0.39	0.48	0.40	0.50	0.61
0.58	0.39	0.68	0.47	0.70	0.52	0.45	0.59	0.57	0.49	0.42	0.51	0.59	0.63	0.68
0.61	0.70	0.63	0.59	0.67	0.66	0.81	0.75	0.51	0.66	0.68	0.55	0.74	0.56	0.65
0.64	0.66	0.57	0.56	0.62	0.72	0.00	0.00	0.00	0.00	0.00	0.54	0.55	0.53	0.52
0.54	0.55	0.53	0.54	0.52	0.54	0.52	0.54	0.54	0.55	0.53	0.51	0.52	0.53	0.53
0.54	0.56	0.56	0.56	0.56	0.56	0.57	0.55	0.56	0.58	0.58	0.59	0.60	0.60	0.61

(These data sets can also be found in the Appendix as Series 2, and can be seen more clearly in Figure 3.1).

After the initial 75 data points, X_t is then taken as following a linearly increasing and decreasing pattern for 66 further points up to $t = 141$. Y_t is simulated from this X_t data using the parameter values obtained after forecasting the West and Harrison data (whilst employing discounting); i.e. using the posterior $\mathbf{m}_{75}$. In order for this simulated Y_t series to be usefully forecastable, we must introduce a random error into the data (or else $f_t = Y_t/n_t$ always). So we take

$$\begin{aligned} Y_t &= n_t \mathbf{F}_t^T \mathbf{m}_t \\ \text{for } \mathbf{m}_t &= \mathbf{g}_t(\mathbf{m}_{t-1}) + \mathbf{w}_t \\ \text{where } \mathbf{w}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t) \,, \end{aligned}$$

(n_t, the sample size, is 66 throughout). $\mathbf{F}_t$, $\mathbf{g}_t(.)$ are as in the model given in section 2.5 and in West and Harrison, and $\mathbf{w}_t$ is simulated from the given

zero-mean Normal distribution, with variance $\mathbf{W}_t$ taken (for convenience of simulation) as being diagonal with constant and equal elements ($\mathbf{W}_t = \sigma^2\mathbf{I}$). Finally, from $t = 142$ onwards, we take random TVR levels by drawing X_t from a uniform distribution on (2,10), and simulate a further 39 data points using the posterior $\mathbf{m}_{141}$, but with the value of κ_{141} shifted slightly from 0.0224 to 0.03. At the beginning of this period, $t = 142$ to 146, we introduce 5 missing data points, and then forecast using $\mathbf{m}_{141}$ as before. The simulated data over this final period is left unchanged in order not to mask the slight shift in behaviour, and throughout the analysis a single discount factor $\delta = 0.97$ is used (the value chosen by West and Harrison).

Before discussing the results of the forecasting, it is worth noting some features of the data. Firstly, it must be remembered that although the TVR levels used between $t = 76$ and $t = 141$ appear rather artificial, the simulated data series Y_t is nevertheless perfectly realistic and could just as easily have been produced by much lower and more random TVR values during a more efficient advertising campaign (i.e. one for which κ_t was generally higher). There is nothing unusual about the X_t values after $t = 141$; in fact they are quite in keeping with TVR patterns and levels seen in a similar analysis from West, Harrison and Migon [37]. Secondly, the shift in κ_t for this last period of data is also very much in keeping with the behaviour of this parameter - changes in the nature of advertising campaigns often produce much larger displacements in κ_t due to changes in the effectiveness of the new campaign. Finally, the missing data values are also a feature of this kind of analysis, and there is nothing unusual about their coming at the start of a new campaign. (The reader is

referred to West, Harrison and Migon for examples of these features). In brief, we are dealing with a data set whose nature is reasonably akin to other data sets of this kind.

The full forecasting run, together with the X_t values, is shown in figure 3.1. The performance is fine up to $t = 141$ and the missing data; however, after $t = 146$ at all but one point the one-step ahead forecasts f_t are below the true data value, and seriously so for the first 15 points. There is only a very slow improvement in performance through to $t = 180$. (This is more clearly seen in Figure 3.2).

Figure 3.1: forecasts of positive respondents, Y (*) using discounting (-) and additive (..) approaches

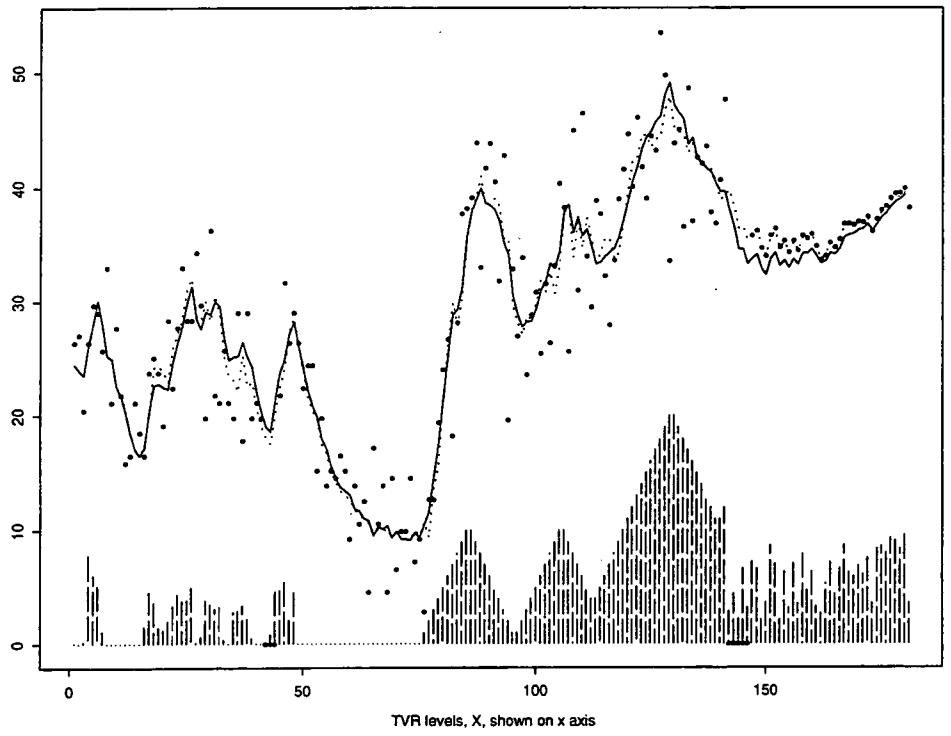
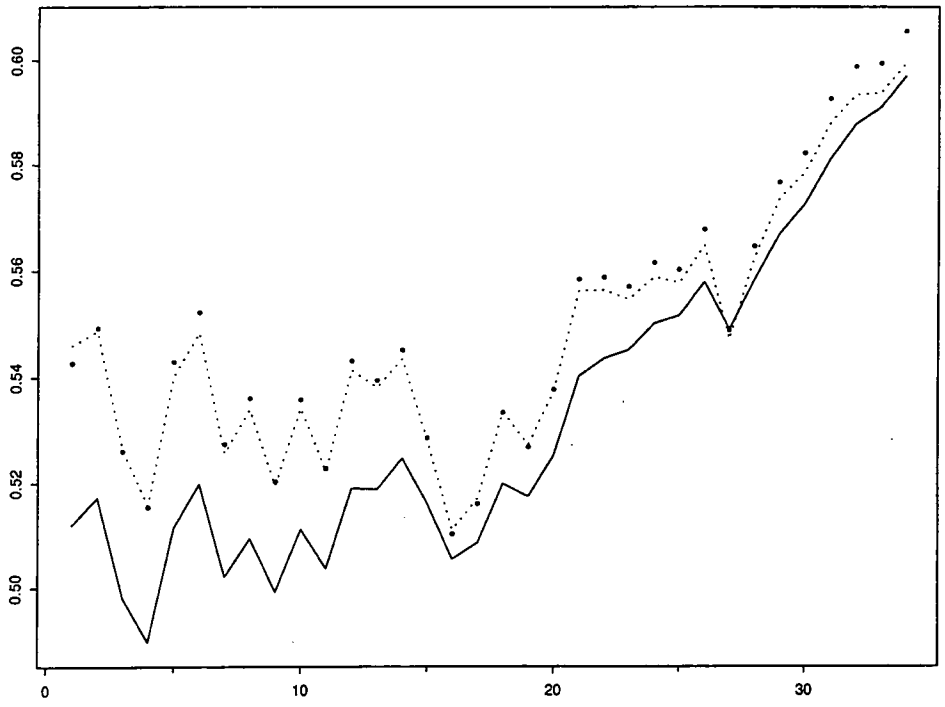


Figure 3.2: forecasts of final 34 Y/n values (*); using discounting (-) and additive (..) approaches



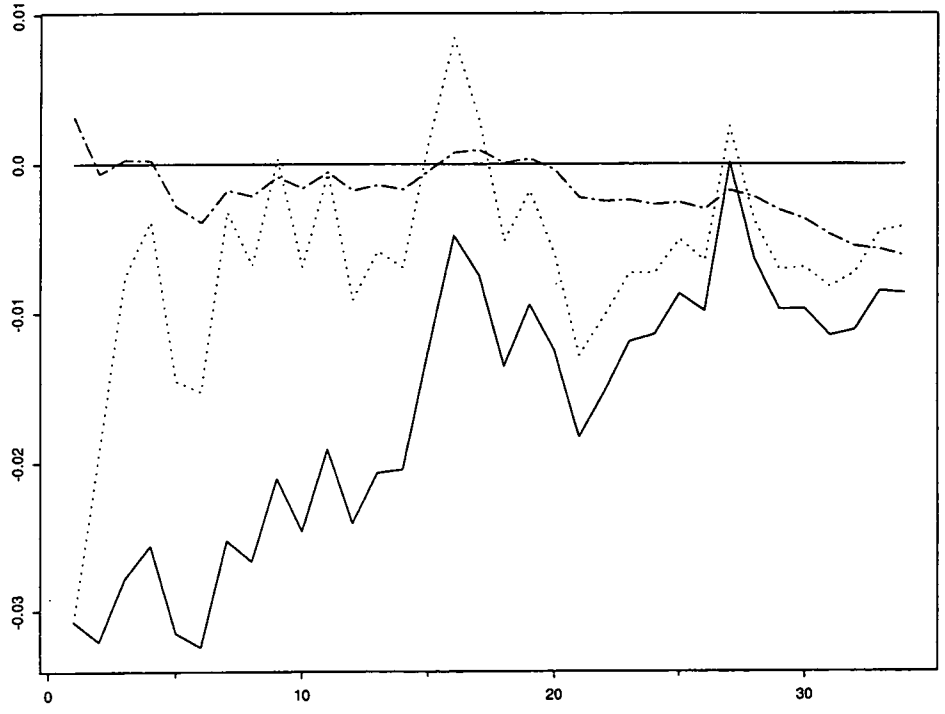
Clearly, we have a problem. If we look at our uncertainties at $t = 141$, we can see the source of it: these uncertainties are given as

$$\mathbf{C}_{141} = \begin{pmatrix} 1.68\times10^{-3} & 4.65\times10^{-4} & -8.35\times10^{-4} & 3.14\times10^{-5} & -1.74\times10^{-3} \\ 4.65\times10^{-4} & 1.57\times10^{-3} & -5.10\times10^{-4} & -3.55\times10^{-5} & 6.09\times10^{-5} \\ -8.35\times10^{-4} & -5.10\times10^{-4} & 8.17\times10^{-4} & -5.68\times10^{-5} & 9.07\times10^{-4} \\ 3.14\times10^{-5} & -3.55\times10^{-5} & -5.68\times10^{-5} & 1.03\times10^{-5} & -6.39\times10^{-5} \\ -1.74\times10^{-3} & 6.09\times10^{-5} & 9.07\times10^{-4} & -6.39\times10^{-5} & 2.10\times10^{-3} \end{pmatrix}$$

Here we see, certainly with respect to the third and fourth parameters, very small (relative to their scale of measurement) values on the diagonal of $\mathbf{C}_{141}$. In the case of κ_t, this value of $\mathbf{C}_{141}[4,4]$ places a ± 2 s.d. level of uncertainty in κ_{141} equal to ± 0.0064. Since $\kappa_{141} = 0.0224$, this places the 'true' value of κ_{141} (namely 0.03, which was the shifted value that the subsequent data was simulated from) outside this range. What we have here is an example of the frailty of the discounting approach, as it has allowed us to express uncertainties

in the state vector that are clearly unrealistic.

Figure 3.3: last 34 residuals from Y/n data; using discounting (-), with intervention (..), and additive (_._) forecasting



We can attempt to improve forecasting performance whilst discounting, by employing the 'safety belt' notion of intervention. The shift in κ_{141} represents the possible effect of a new advertising campaign, as already mentioned; if we were able to anticipate a potential change in the nature of the model such as this,

the intervention procedure has us decrease the discounting factor at the relevant time point to a much smaller value, hence producing a marked increase in our uncertainties within $\mathbf{R}_t$ and $\mathbf{C}_t$. This "enables the model to adapt rapidly to any changes in the parameters .... at these times" (West, Harrison and Migon). The results, for the last 34 points, of following this method by setting $\delta = 0.1$ at $t = 147$ is shown in Figure 3.3 above. (Note that this figure, along with ensuing discussions, utilises the residuals defined by $f_t - (Y_t/n_t)$ - the forecast minus the data point.)

Despite the obvious improvement in absolute size (the variance is nearly halved), the pattern in the residuals remains much the same with predominantly large negative values, which are noticeably more so when X_t suddenly increases, indicating a failure of the model to adapt to the increasing 'penetration' effect of the advertising. And this is despite more than due caution at $t = 147$ (remember that the shift in κ_{141} was relatively small in the context of the changes that *can* occur in new advertising campaigns), where intervention resulted in a tenfold increase in our uncertainties (an increase of over 200% in standard deviation).

The solution to the slowness of the model to adapt to changes, even with intervention, is straightforward. In such a long series, where there is a lot of information potentially available in relation to every parameter in the state vector, we must resort to an additive form for $\mathbf{W}_t$ which contains our subjective judgements on $\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, the 'lower bound' on the uncertainties in $\mathbf{R}_t$ (see section 3.4) which are then transferred to $\mathbf{C}_t$. This will avoid the unrealistic

values we see in some places on the diagonal of, for instance, $\mathbf{C}_{141}$ above.

The choice of values in $\mathbf{W}_t$ are, of course, extremely subjective and require much thought in relation to the scale of the parameters in the state vector. The discussion of the many interpretations involved in this choice of $\mathbf{W}_t$ is left to the next section; for now, it suffices to show how the simple conclusions drawn by even an inexperienced practitioner, when attempting to assign values to $\mathbf{W}_t$, can be far more effective than relying on the limiting choice of a single discount factor.

We consider the diagonal of $\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ element-wise. From section 3.4 we learn that we are trying to assign an uncertainty to a parameter value at time t, *given* its value at time $t-1$. The conclusions are these:

$$\begin{aligned}
\mathrm{Var}(\alpha_t|\alpha_{t-1}) &= 2.5\times10^{-5} \\
\mathrm{Var}(\beta_t|\beta_{t-1}) &= 1\times10^{-4} \\
\mathrm{Var}(\rho_t|\rho_{t-1}) &= 5\times10^{-5} \\
\mathrm{Var}(\kappa_t|\kappa_{t-1}) &= 2.5\times10^{-5}
\end{aligned}$$

These variances are based on taking the measurement scales of the given parameters, on which to base beliefs in their one-step ahead variation, as being the same as in the prior $\mathbf{m}_0$ used by West and Harrison, i.e. $\mathbf{m}_0 = (0.10, 0.85, 0.90, 0.02, 0.30)^T$. The variance of the fifth parameter, E_t, given E_{t-1}, is more complicated. This TVR effect is given as $E_t = (\beta_t - \alpha_t) - (\beta_t - \alpha_t - \rho_t E_{t-1})e^{-\kappa_t X_t}$, and as such its variance given E_{t-1} will depend not only upon the above variances for the other parameters, but also upon the very

value of E_{t-1}. We can, however, throw light on the problem by considering the (important) special case of $X_t = 0$ - then $E_t = \rho_t E_{t-1}$, and so conditioning on $\boldsymbol{\theta}_{t-1}$ gives

$$\begin{aligned} \text{Var}(E_t|E_{t-1}) &= E_{t-1}^2 \text{Var}(\rho_t|\rho_{t-1}) \\ &= 5\times10^{-5} \times E_{t-1}^2 \ . \end{aligned}$$

Typically, E_{t-1} is around 0.4; this leads to

$$\text{Var}(E_t|E_{t-1}) = 8\times10^{-6} \ .$$

This is an approximate lower bound for $\text{Var}(E_t|E_{t-1})$ - and as we are attempting to assign lower bounds to $\text{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, it seems sensible to use this as our final choice. The analysis is usefully simplified by evaluating the rest of $\mathbf{W}_t$ (namely the covariance terms) through discounting using the same discount factor of 0.97 as before. Hence we still evaluate $\mathbf{W}_t$ as $\frac{\delta}{1-\delta}(\mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T)$, but importantly with the diagonal elements replaced as above.

The *entire* data set Y_t is then forecasted using X_t exactly as before. It should be noted that the shift at $t = 141$ in κ_t was to the mean value of κ_{141} in $\mathbf{m}_{141}$ from the two separate analyses using discounting and the additive form for $\mathbf{W}_t$; for reference, these two posteriors are

$$\mathbf{m}_{141} = (0.125, 0.763, 0.902, 0.0224, 0.487)^T \quad \text{from the discounting method ,}$$

and $\mathbf{m}_{141} = (0.0895, 0.702, 0.915, 0.0367, 0.548)^T$ from using the above additive $\mathbf{W}_t$.

Hence taking $\kappa_{141} = 0.030$ as the shifted value (the mean of κ_{141} from the two approaches, to 3 d.p.) does not 'favour' either method, and, indeed, simulating the data using the discounting posterior $\mathbf{m}_{141}$ results in *all* the parameters being shifted by varying amounts in $\mathbf{m}_{141}$ from the additive approach, over the whole of this final period.

The results of forecasting using this additive $\mathbf{W}_t$ are shown in Figures 3.1, 3.2 and 3.3. The improvement is remarkable.

Finally, for comparison, the residual means and variances from the last 34 data points, for all 3 forecasting runs, are given in Table 3.2.

Table 3.2: Residual means and variances from $t = 147$ to $t = 180$.

Forecasting method	Residual mean	Residual variance
Discounting, no intervention	-0.0166	7.93×10^{-5}
Discounting, intervention	-0.00645	4.74×10^{-5}
Additive $\mathbf{W}_t$	-0.00184	4.10×10^{-6}

3.4 Interpretation and further discussion of discounting

Consider now the following derivation of the prior variance, $\mathbf{R}_t$. From the state equation, 2.2,

$$\mathrm{E}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}] = \mathrm{E}[(\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t)|\boldsymbol{\theta}_{t-1}] = \mathbf{G}_t\boldsymbol{\theta}_{t-1} \ ,$$

$$\text{so} \ \ \mathrm{Var}(\mathrm{E}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}]) = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T \ ,$$

and also

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \mathrm{Var}((\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t)|\boldsymbol{\theta}_{t-1}) \\
&= \mathrm{Var}(\mathbf{G}_t\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_{t-1}) + \mathbf{W}_t \quad (\text{since } \mathbf{w}_t \text{ is independent of } \boldsymbol{\theta}_{t-1}) \\
&= \mathbf{W}_t \, .
\end{aligned}$$

So since

$$\mathrm{Var}(\boldsymbol{\theta}_t|D_{t-1}) = \mathrm{Var}(\mathrm{E}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}]|D_{t-1}) + \mathrm{E}[\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})|D_{t-1}] \, ,$$

we have

$$\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{W}_t \, , \quad \text{as in section 2.2.}$$

Now we have $\mathbf{W}_t$ expressed in a more readily interpretable form - and one which we made use of in the previous section - as the quantity $\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$. The values in $\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ are a reflection of our uncertainty in the state at time t, when we *know* what our state is at time $t-1$. This can be thought of as a 'lower limit' of the values in $\mathbf{R}_t$, so that it is impossible to be in the position of having more prior information at time t about our state than is realistic within our model - something we have already seen in section 3.3, where discounting led to negligible absolute values within our state vector uncertainties.

This natural interpretation of $\mathbf{W}_t$ helped to negotiate the potentially disastrous case of small $\mathbf{C}_t$ values earlier. The concept of discounting is much harder to interpret, however - it arises largely from consideration of the closed, steady DLM (see section 4.3). This model type produces a (generally) fast convergence

of $C_t \to C$, and hence $R_t \to R = C+W$, $Q_t \to Q = R+V$, and $A_t \to A = R/Q$. Thus

$$Q = \frac{R}{A} = R + V, \ \Rightarrow R = \frac{VA}{1-A} ,$$
$$\text{and since } C = \frac{R}{Q}V = AV, \ R = \frac{C}{1-A} .$$

This leads to the conception of the discounting approach, since $R = C + W$ implies that $W = (\frac{A}{1-A})C$, and then we take $\delta = 1 - A$. However, the methodology crucially depends upon the specification of the observational variance V_t, and also having scalar $\mathbf{F} = \mathbf{G} = 1$ - we cannot derive $\mathbf{R} = \frac{\mathbf{C}}{\mathbf{1-A}}$ even in the general constant TSDLMs of section 2.7 where $\mathbf{F}, \mathbf{G}$ are not the identity, let alone in the above non-linear example where we do not even specify a V_t. Indeed, in many non-linear models, there is no such specification (variation in the data Y_t is transferred instead via the conjugate prior analysis) and so discounting has no convenient interpretation or derivation. In fact, the general discounting statement, that W_t is a fixed percentage of the posterior state variance at time $t-1$, can arise from the peculiar position of requiring the uncertainty in our state vector at time t, given the state at time $t-1$, to be proportional to $\boldsymbol{\theta}_{t-1}\boldsymbol{\theta}_{t-1}^T$ (up to a constant). For if we take $\text{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \mathbf{A}\boldsymbol{\theta}_{t-1}\boldsymbol{\theta}_{t-1}^T + \mathbf{B}$, for $\mathbf{A}$ and $\mathbf{B}$ matrix constants ($\mathbf{A}$ diagonal), we find

$$\mathbf{W}_t = \text{E}[\text{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})] = \mathbf{A}(\text{Var}(\boldsymbol{\theta}_{t-1}) + \mathbf{m}_{t-1}\mathbf{m}_{t-1}^T) + \mathbf{B} ;$$

thus taking $\mathbf{B} = -\mathbf{A}\mathbf{m}_{t-1}\mathbf{m}_{t-1}^T$ gives

$$\mathbf{W}_t = \mathbf{A}\mathbf{C}_{t-1} \ ,$$

which is (effectively) our discounting statement. Thus it is hard to credit an inexperienced practitioner with having a 'natural feel' for a concept that is not naturally interpretable in many examples of use - this comment by Ameen and Harrison is instead, one feels, a reflection of the parallel comment that discounting is parsimonious and easy to apply; however, we have just seen what practical difficulties this can lead us into without due care.

Finally, it is wise to examine Ameen and Harrison's statement that "once the discount factors are chosen, established methods for the on-line estimation of the observational variance may be applied" a little more closely. The 'established methods' include specifying a joint Normal/Gamma distribution for Y_t and $\phi = 1/V$ (constant observation variance). By specifying a prior Gamma distribution for ϕ, $\phi \sim \Gamma(n_{t-1}, d_{t-1})$, and following a full conjugate analysis using $p(\phi|D_t) \ \alpha \ p(Y_t|\phi)p(\phi|D_{t-1})$, we are led to expressing a fully defined and parameterised posterior for ϕ, namely $\phi \sim \Gamma(n_{t-1} + 1/2, d_{t-1} + \frac{1}{2}(Y_t - f_t)^2/Q_t^*)$, (where $\frac{Q_t^*}{\phi} = Q_t$), which allows us to update our on-line estimate of $V = 1/\phi$. However, this conjugate prior analysis (and, indeed, other 'established methods') has absolutely no reliance on the choice of discount factors - it merely requires that we specify a prior mean and variance for $\boldsymbol{\theta}_t$ (and hence Y_t) which is passed through the analysis via $\mathbf{R}_t$ whether we use discounting *or* an additive $\mathbf{W}_t$. This issue, as when looking at the motivation behind discounting, is not

even relevant in the advertising awareness example, or in any other non-linear models, where an observational variance is not specified. (The above conjugate prior analysis should be treated with caution anyhow; our hands are tied in belief expressions about Y_t, since specifying a joint Normal/Gamma distribution for Y_t and $\phi = 1/V$ no longer allows us to make independent statements about the mean and variance of $(Y_t|\phi) \sim \mathrm{N}(f_t, Q_t^*/\phi))$.

3.5 The step-ahead forecast error distribution as a method for estimating $\mathbf{W}_t$

It is all very well to criticise the method of discounting, but we must remember that its real strength lies in its ease of application compared with the enormity of estimating the matrix $\mathbf{W}_t$ (as well as the (generally) scalar V_t), which we have seen to be crucial specifications in the DLM framework, and whose estimation we still have little help with.

In choosing $\mathbf{W}_t$ we are still faced with various problems. Firstly, although interpreting the matrix as $\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ assists us in our specification of the diagonal elements of $\mathbf{W}_t$, we are not furthered in the search for the off-diagonal, covariance terms. These require tackling the awkward questions of how one parameter within the state vector is influenced by the perturbations in another, and there are some $p(p-1)/2$ (where $\boldsymbol{\theta}_t$ is p-dimensional) of these questions to answer. Very often we cannot realistically get further than place a sign on these covariances. Secondly, even the most experienced of practitioners will still be somewhat worried at accepting blindly that their final evaluation of $\mathbf{W}_t$ will

serve them reliably throughout the analysis.

The potential solution to our worries lies in their very wording - "...serve them reliably *throughout* the analysis". If it was possible to have a repeatedly updated on-line estimation of $\mathbf{W}_t$, obtained through feedback from each time point of the analysis, then most practitioners would be greatly relieved - the system evolution variance could be happily left to 'update itself'. Here we develop a method of looking at covariances of $(Y_{t+j}, Y_{t+k}|D_t)$, for $k > j$, that at each stage produces a simple equation in a linear combination of the unknown elements of W_t.

From the defining equations 2.1 and 2.2, we have that

$$Y_{t+1} = \mathbf{F}_{t+1}^T\boldsymbol{\theta}_{t+1} + v_{t+1} = \mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\boldsymbol{\theta}_t + \mathbf{F}_{t+1}^T\mathbf{w}_{t+1} + v_{t+1}\ ;$$

$$\begin{aligned} Y_{t+2} = \mathbf{F}_{t+2}^T(\mathbf{G}_{t+2}\boldsymbol{\theta}_{t+1}+\mathbf{w}_{t+2})+v_{t+2} &= \mathbf{F}_{t+2}^T(\mathbf{G}_{t+2}(\mathbf{G}_{t+1}\boldsymbol{\theta}_t + \mathbf{w}_{t+1}) + \mathbf{w}_{t+2}) + v_{t+2} \\ &= \mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{G}_{t+1}\boldsymbol{\theta}_t+\mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{w}_{t+1} \\ &\quad + \mathbf{F}_{t+2}^T\mathbf{w}_{t+2} + v_{t+2}\ ; \end{aligned}$$

and, generally,

$$\begin{aligned} Y_{t+k} &= \mathbf{F}_{t+k}^T\mathbf{G}_{t+k}\mathbf{G}_{t+k-1}\ldots\mathbf{G}_{t+1}\boldsymbol{\theta}_t + \mathbf{F}_{t+k}^T\mathbf{G}_{t+k}\ldots\mathbf{G}_{t+2}\mathbf{w}_{t+1} + \ldots \\ &+ \mathbf{F}_{t+k}^T\mathbf{G}_{t+k}\ldots\mathbf{G}_{t+j+1}\mathbf{w}_{t+j} + \ldots + \mathbf{F}_{t+k}^T\mathbf{w}_{t+k} + v_{t+k}. \end{aligned} \tag{3.1}$$

Thus, given the posterior $(\boldsymbol{\theta}_t|D_t) \sim (\mathbf{m}_t, \mathbf{C}_t)$ at time t, we can calculate

$$\mathrm{E}[\mathrm{Y}_{t+1}|D_t] = \mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\mathbf{m}_t\ ,$$

$$\mathrm{E}[\mathrm{Y}_{t+2}|D_t] = \mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{G}_{t+1}\mathbf{m}_t \ ,$$

$$\text{and generally } \mathrm{E}[\mathrm{Y}_{t+k}|D_t] = \mathbf{F}_{t+k}^T\mathbf{G}_{t+k}\mathbf{G}_{t+k-1}\ldots\mathbf{G}_{t+1}\mathbf{m}_t. \qquad (3.2)$$

(For the rest of the section, all expectations are with reference to time t, and so the conditioning upon D_t is to be assumed). Additionally, we have that

$$\mathrm{E}[\mathrm{Y}_{t+2}\mathrm{Y}_{t+1}] = \mathrm{E}[(\mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{G}_{t+1}\boldsymbol{\theta}_t)(\mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\boldsymbol{\theta}_t)] + \mathrm{E}[(\mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{w}_{t+1})(\mathbf{F}_{t+1}^T\mathbf{w}_{t+1})]$$

since all other pairs in the product are independent, with $\mathrm{E}[\mathbf{w}_{t+1}] = \mathrm{E}[\mathbf{w}_{t+2}]$ $(= \mathrm{E}[v_{t+1}] = \mathrm{E}[v_{t+2}]) = \mathbf{0}$ (or 0). Now, since each expectation is a scalar, we can take transposes of the right-hand half of each to get

$$\begin{aligned}\mathrm{E}[\mathrm{Y}_{t+2}\mathrm{Y}_{t+1}] &= \mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{G}_{t+1}\mathrm{E}[\boldsymbol{\theta}_t\boldsymbol{\theta}_t^T]\mathbf{G}_{t+1}^T\mathbf{F}_{t+1} + \mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathrm{E}[\mathbf{w}_{t+1}\mathbf{w}_{t+1}^T]\mathbf{F}_{t+1} \\ &= \mathbf{F}_{t+2}^T\mathbf{G}_{t+2}(\mathbf{G}_{t+1}(\mathbf{C}_t + (\mathbf{m}_t\mathbf{m}_t^T))\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{F}_{t+1} \ ,\end{aligned}$$

from noting that $\mathrm{Var}(\boldsymbol{\theta}_t|D_t) = \mathbf{C}_t = \mathrm{E}[(\boldsymbol{\theta}_t|D_t)(\boldsymbol{\theta}_t|D_t)^T] - (\mathrm{E}[\boldsymbol{\theta}_t|D_t])(\mathrm{E}[\boldsymbol{\theta}_t|D_t])^T = \mathrm{E}[\boldsymbol{\theta}_t\boldsymbol{\theta}_t^T] - (\mathbf{m}_t\mathbf{m}_t^T)$, and that $\mathrm{Var}(\mathbf{w}_{t+1}) = \mathrm{E}[\mathbf{w}_{t+1}\mathbf{w}_{t+1}^T] = \mathbf{W}_{t+1}$. Then, after denoting the k^{th} step-ahead forecast 3.2 by $f_{t+k,k}$, we can evaluate

$$\mathrm{E}[(Y_{t+2} - f_{t+2,2})(Y_{t+1} - f_{t+1})] = \mathrm{E}[Y_{t+2}Y_{t+1}] - f_{t+2,2}\mathrm{E}[Y_{t+1}] - f_{t+1}\mathrm{E}[Y_{t+2}] + f_{t+2,2}f_{t+1} \ ;$$

the last two terms cancel and we are left with

$$\begin{aligned}\mathrm{E}[(Y_{t+2} - f_{t+2,2})(Y_{t+1} - f_{t+1})] &= \mathbf{F}_{t+2}^T\mathbf{G}_{t+2}(\mathbf{G}_{t+1}(\mathbf{C}_t + \mathbf{m}_t\mathbf{m}_t^T)\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{F}_{t+1} \\ &\quad - (\mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{G}_{t+1}\mathbf{m}_t)(\mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\mathbf{m}_t)\end{aligned}$$

$$= \mathbf{F}_{t+2}^T \mathbf{G}_{t+2}(\mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{F}_{t+1} \ , \qquad (3.3)$$

since the $\mathbf{m}_t\mathbf{m}_t^T$ terms cancel through taking the transpose of $(\mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\mathbf{m}_t)$. This elimination of the $\mathbf{m}_t\mathbf{m}_t^T$ term is why, incidentally, it is computationally wise to look at the covariances of the step-ahead forecasts, as opposed to simply the expectation of their products. (Note that $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t) \equiv \mathrm{E}[(Y_{t+2} - f_{t+2,2})(Y_{t+1} - f_{t+1})]$).

Given that at all time points t we know $\mathbf{C}_t$ exactly, what this represents is simply a linear combination of the elements of $\mathbf{W}_{t+1}$, and so is just one equation in $p(p+1)/2$ (potential) unknowns: we need to generate more equations. Before moving further time steps ahead, we should examine the above expectation again, only this time replacing $f_{t+2,2}$ with $f_{t+2,1}$ (effectively $\mathrm{E}[Y_{t+2}|D_{t+1}]$). Unfortunately, since $\mathrm{E}[(Y_{t+1} - f_{t+1})|D_t] = 0$, this quickly reduces to equation 3.3; hence we must look farther afield... With reference to equation 3.1 above, we find

$$\begin{aligned}
\mathrm{E}[Y_{t+3}Y_{t+1}] &= \mathrm{E}[(\mathbf{F}_{t+3}^T\mathbf{G}_{t+3}\mathbf{G}_{t+2}\mathbf{G}_{t+1}\boldsymbol{\theta}_t)(\mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\boldsymbol{\theta}_t)] \\
&+ \mathrm{E}[\mathbf{F}_{t+3}^T\mathbf{G}_{t+3}\mathbf{G}_{t+2}\mathbf{w}_{t+1})(\mathbf{F}_{t+1}^T\mathbf{w}_{t+1})] \ , \\
&\quad \text{(since all other covariances are again 0)}, \\
&= \mathbf{F}_{t+3}^T\mathbf{G}_{t+3}\mathbf{G}_{t+2}(\mathbf{G}_{t+1}(\mathbf{C}_t + \mathbf{m}_t\mathbf{m}_t^T)\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{F}_{t+1} \ ,
\end{aligned}$$

and are led to

$$\mathrm{E}[(Y_{t+3} - f_{t+3,3})(Y_{t+1} - f_{t+1})] = \mathrm{E}[Y_{t+3}Y_{t+1}] - f_{t+3,3}f_{t+1}$$

$$= \mathbf{F}_{t+3}^T \mathbf{G}_{t+3}\mathbf{G}_{t+2}(\mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{F}_{t+1} \ .$$

It is now easy to prove, using induction via the observation and state equations together with 3.2, the more general result of

$$\mathrm{E}[(Y_{t+k} - f_{t+k,k})(Y_{t+1} - f_{t+1})] = \mathbf{F}_{t+k}^T \mathbf{G}_{t+k}\mathbf{G}_{t+k-1}\ldots\mathbf{G}_{t+2}(\mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{F}_{t+1}. \tag{3.4}$$

However, to obtain enough information to solve for $p(p+1)/2$ unknowns in $\mathbf{W}_t$ using solely equation 3.4, we would have to wait until at least time $(p(p+1)/2)+1$ into the analysis. For the advertising awareness model with $p = 5$ this represents only a time lag of at most 16 steps; for a more complicated seasonal model with monthly seasonality components (depending upon how the practitioner views the covariance terms, of course - perturbations in the $(j+6)^{\text{th}}$ component may well be deemed unlikely to affect the j^{th} term) we could be facing a wait until more than 50 steps into the analysis. But help is at hand from looking at covariances between the j^{th} and k^{th} step-ahead forecast errors, for both $j,\ k > 1$. For instance, from 3.1 above, the covariance between the 2nd and 3rd steps-ahead forecast errors is found from

$$\begin{aligned}
\mathrm{E}[Y_{t+3}Y_{t+2}] &= \mathrm{E}[(\mathbf{F}_{t+3}^T(\mathbf{G}_{t+3}(\mathbf{G}_{t+2}(\mathbf{G}_{t+1}\boldsymbol{\theta}_t + \mathbf{w}_{t+1}) + \mathbf{w}_{t+2}) + \mathbf{w}_{t+3})) \\
&\qquad (\mathbf{F}_{t+2}^T(\mathbf{G}_{t+2}(\mathbf{G}_{t+1}\boldsymbol{\theta}_t + \mathbf{w}_{t+1}) + \mathbf{w}_{t+2}))] \\
&= \mathrm{E}[(\mathbf{F}_{t+3}^T\mathbf{G}_{t+3}\mathbf{G}_{t+2}\mathbf{G}_{t+1}\boldsymbol{\theta}_t)(\mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{G}_{t+1}\boldsymbol{\theta}_t)] \\
&+ \mathrm{E}[(\mathbf{F}_{t+3}^T\mathbf{G}_{t+3}\mathbf{G}_{t+2}\mathbf{w}_{t+1})(\mathbf{F}_{t+2}^T\mathbf{G}_{t+2}\mathbf{w}_{t+1})]
\end{aligned}$$

$$+ \mathrm{E}[(\mathbf{F}_{t+3}^T \mathbf{G}_{t+3} \mathbf{w}_{t+2})(\mathbf{F}_{t+2}^T \mathbf{w}_{t+2})] \quad \text{(again, all other covariances are zero)}$$

$$= \mathbf{F}_{t+3}^T \mathbf{G}_{t+3}((\mathbf{G}_{t+2}(\mathbf{G}_{t+1}(\mathbf{C}_t + \mathbf{m}_t\mathbf{m}_t^T)\mathbf{G}_{t+1}^T + \mathbf{W}_{t+1})\mathbf{G}_{t+2}^T) + \mathbf{W}_{t+2})\mathbf{F}_{t+2} \, .$$

Note that this time we have introduced a $\mathbf{W}_{t+2}$ term. In order that we should be able to solve this equation for $\mathbf{W}_{t+1}$ we must now assume not only that the $\mathbf{w}_{t+j}$ terms are all independent, but that they are identically distributed as well - i.e. we are looking for $\mathbf{W}_t = \mathbf{W}$, a constant. This assumption is nothing more than a statement of practitional fact, as once a value of $\mathbf{W}_t$ is finalised (or, for that matter, once discount parameters are decided) it will be held constant until there is some cause for intervention. Moreover, we have seen that constant TSDLMs are vital models in many applications.

Having made the assumption that $\mathbf{W}_t = \mathbf{W}$ is constant, and following our previous methodology, we find the covariance term

$$\mathrm{E}[(Y_{t+3} - f_{t+3,3})(Y_{t+2} - f_{t+2,2})] = \mathrm{E}[Y_{t+3}Y_{t+2}] - f_{t+3,3}f_{t+2,2}$$

$$= \mathbf{F}_{t+3}^T \mathbf{G}_{t+3}((\mathbf{G}_{t+2}(\mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T + \mathbf{W})\mathbf{G}_{t+2}^T) + \mathbf{W})\mathbf{F}_{t+2} \, .$$

Again, it is easy to generalise this from equations 3.1 and 3.2, giving

$$\mathrm{E}[(Y_{t+k} - f_{t+k,k})(Y_{t+j} - f_{t+j,j})] = \mathbf{F}_{t+k}^T \mathbf{G}_{t+k} \ldots \mathbf{G}_{t+j+1}(\mathbf{G}_{t+j} \ldots \mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T \ldots \mathbf{G}_{t+j}^T$$

$$+ \mathbf{G}_{t+j} \ldots \mathbf{G}_{t+2}\mathbf{W}\mathbf{G}_{t+2}^T \ldots \mathbf{G}_{t+j}^T + \ldots$$

$$\ldots + \mathbf{G}_{t+j}\mathbf{W}\mathbf{G}_{t+j}^T + \mathbf{W})\mathbf{F}_{t+j} \quad (\text{for } k > j). \tag{3.5}$$

In theory, therefore, we have some $i(i-1)/2$ equations resulting from looking

at all possible covariances between the j^{th} and k^{th} step-ahead forecast errors, for $1 \leq j < k \leq i$. But we can also look at these covariances for $j = k$, or, in other words, from $\text{Var}(Y_{t+k}|D_t)$. Now we have, simply, that

$$\text{Var}(Y_{t+1}|D_t) = Q_{t+1} = \mathbf{F}_{t+1}^T(\mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T + \mathbf{W})\mathbf{F}_{t+1} + V \ ,$$

and generally,

$$\begin{aligned}\text{Var}(Y_{t+k}|D_t) = \mathbf{F}_{t+k}^T(\mathbf{G}_{t+k}\ldots\mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}_{t+1}^T\ldots\mathbf{G}_{t+k}^T + \mathbf{G}_{t+k}\ldots\mathbf{G}_{t+2}\mathbf{W}\mathbf{G}_{t+2}^T\mathbf{G}_{t+k}^T + \ldots \\ \ldots + \mathbf{G}_{t+k}\mathbf{W}\mathbf{G}_{t+k}^T + \mathbf{W})\mathbf{F}_{t+k} + V \ . \qquad (3.6)\end{aligned}$$

Hence, after i time points, we can generate some $i(i+1)/2$ equations in the $p(p+1)/2$ unknowns of $\mathbf{W}$ as well as the extra unknown of V. Thus for a p-dimensional state vector we need only wait until the $(p+1)$st step in the analysis to be able to obtain enough information about $\mathbf{W}$ and V to solve for them, theoretically, and thus start an on-line updating estimation procedure. In practice, however, not all of the $i(i+1)/2$ equations will be linearly independent, and we may well be forced to wait until further into the analysis before we generate a sufficient set of linearly independent equations (presuming that this is feasible at all; section 3.6 deals with the potential non-uniqueness of $\mathbf{W}$).

At this point it should be noted that there is nothing new in wishing to find an on-line estimation procedure for V_t and $\mathbf{W}_t$; neither is there anything particularly original in commenting that covariances between step-ahead forecasts

will produce equations in the elements of $\mathbf{W}_t$. Indeed, the search for identification methods of the noise variances within the Kalman Filter was underway even before Harrison and Stevens' [18] first (1971) paper on the DLM, in two papers by Mehra (1970 [28], and then 1972 [29]) in the IEEE Transactions on Automatic Control, and one by Godbole [13] which extended Mehra's method in 1974 in the same Journal. Mehra proposed various sub-optimal filtering identification methods, from finding maximum likelihood estimates of both Q_t and the adaptive coefficient $\mathbf{A}_t$ (which can then be solved to give estimates of both V_t and $\mathbf{W}_t$ through invertion of the relations of the Kalman Filter), to looking at the sequence of output correlations within the data series $\{Y_t\}$. Godbole extended the applicability of these methods by noting that they do not rely on *a priori* knowledge of the mean of the noise sequences v_t and $\mathbf{w}_t$ (a critical assumption made by Mehra), and also allowed correlations to exist between the two noise sequences.

However, the major problem with these procedures, together with our proposed solution of equations 3.5 and 3.6, is that they all run the not-inconsiderable risk of divergence of their on-line estimates, since they all involve some form of sub-optimal filtering. This is considered at length at the beginning of chapter 4. Further to this major practitional obstacle, all the other procedures referenced above generate a severely restricted number of equations in the unknown elements of $\mathbf{W}_t$, and may hence require the practitioner to 'put additional restrictions' on this matrix, if it is indeed not uniquely determinable from the equations available. In fact, the only illustrative example given anywhere of the applicability of any of these complex sub-optimal filtering procedures is in

Mehra [28], where severe restrictions are placed upon both $\mathbf{V}_t$ and $\mathbf{W}_t$ due to the computational complexity and potential divergence problems of the method (a bivariate data series Y_t is assumed, giving an observational 2×1 noise vector $\mathbf{v}_t$, with associated 2×2 variance matrix $\mathbf{V}_t$).

Some of these ideas and associated problems then resurfaced in 1980 in a paper by Lee [25], who merely simplifies the unnecessarily intricate calculations of Mehra and Godbole by adopting the slightly different approach of constructing the minimal polynomial of $\mathbf{G}_t = \mathbf{G}$, the constant state evolution matrix, and then using this minimal polynomial to define a sequence z_t as a linear combination of the data points $Y_t, \ldots, Y_{t-m}$, where the minimal polynomial is of degree m. Calculating covariances of z_t, z_{t-i}, $0 \leq i \leq m$, as opposed to covariances between Y_{t+j} and Y_{t+k}, does become simpler computationally - however, these covariances in the sequence z_t are all 0 beyond lag m, and so the number of equations that it is possible to generate from this method is again severely limited, this time to $m + 1$. If the number of unknowns in $\mathbf{W}$ is greater than this, then Lee's advice is also to "put additional restrictions on the form" of this matrix - and as Ameen and Harrison [2] state, "tell that to a practitioner and he is going to get very upset"!

Lee's approach may be cunning from a computational viewpoint, but in constructing the sequence z_t from the minimal polynomial in $\mathbf{G}$ he is losing much possible information about the state evolution variance $\mathbf{W}$. By looking at covariances between the output data points Y_{t+i} directly, we can generate more potentially linearly independent equations in the unknown elements of $\mathbf{W}$ and V. However, our proposed solution method not only also faces the

divergence problems afflicting every other similar method - which, as stated, will be dealt with in chapter 4 - but has only *apparently* generated a set of sufficient equations to solve for $\mathbf{W}$ and V. What we now have to investigate is the non-uniqueness of $\mathbf{W}$.

3.6 Non-uniqueness of W

Whilst we have been quick to draw attention to the perils of discounting, we must be equally critical of our above method - and, indeed, any method that endeavours to solve for $\mathbf{W}$ in a constant TSDLM. There is a fundamental restriction to be appreciated in attempting to define fully an additive system variance matrix.

This section is simply concerned with that deceptively easy-to-use phrase: "define fully". In the literature reviews of the previous section there is constant reference to authors who were aware that their respective approaches may still not have been sufficient to solve uniquely for the $p(p+1)/2$ terms of this matrix, even having generated a potentially sufficient set of equations. Ameen and Harrison, however, were aware of this crucial overparameterisation of $\mathbf{W}$ with respect to the forecast function $f_t(k)$, and it is almost certainly this non-uniqueness that led to the search for an alternative method of specifying the loss of information in the prior variance $\mathbf{R}_t$.

It is best to motivate the rest of this section via an example that appears in both Harrison and Akram [17], and Ameen and Harrison. Consider the model

$$M' = \{\mathbf{E}_2, \mathbf{J}_2(1), V, \mathbf{W}'\} = \left\{(1,0), \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, V, \mathbf{W}' = \begin{pmatrix} W_1 + aW_2 & aW_2 \\ aW_2 & W_2 \end{pmatrix}\right\},$$

where $a(a-1) \leq W_1/W_2$ for $\mathbf{W}'$ to be a valid variance matrix (note that this 2-dimensional constant TSDLM is already in canonical form, with $\mathbf{G}$ having 1 as a repeated eigenvalue of multiplicity 2). Then, suppose we have chosen two distinct specifications of $\mathbf{W}'$ from this model class, the first with some particular $a \neq 0$ in M', the second with $a = 0$ giving

$$M = \left\{(1,0), \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, V, \mathbf{W} = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}\right\}.$$

Given that we are working with the *convergent* form of both models, it is easy to show from Lemma 3.2 (proven later in this section) that in adding $\delta\mathbf{W} = \begin{pmatrix} aW_2 & aW_2 \\ aW_2 & 0 \end{pmatrix}$ to a specified $\mathbf{W}$ we change the convergent value of $\mathbf{C}$ under M' to $\mathbf{C}' = \mathbf{C} + \delta\mathbf{C}$ where $\delta\mathbf{C} = \begin{pmatrix} 0 & 0 \\ 0 & -aW_2 \end{pmatrix}$. Then under M', we have

$$\begin{aligned}
\mathrm{Var}(Y'_{t+1}|D_t) &= Q'_{t+1} = \mathbf{E}_2^T\mathbf{R}'\mathbf{E}_2 + V \\
&= \mathbf{E}_2^T\mathbf{J}_2(1)\mathbf{C}'\mathbf{J}_2(1)^T\mathbf{E}_2 + \mathbf{E}_2^T\mathbf{W}'\mathbf{E}_2 + V \\
&= \mathbf{E}_2^T\mathbf{J}_2(1)\delta\mathbf{C}\mathbf{J}_2(1)^T\mathbf{E}_2 + \mathbf{E}_2^T\delta\mathbf{W}\mathbf{E}_2 + Q_{t+1} \\
&= \mathbf{E}_2^T\mathbf{J}_2(1)\delta\mathbf{C}\mathbf{J}_2(1)^T\mathbf{E}_2 + \mathbf{E}_2^T\delta\mathbf{W}\mathbf{E}_2 + \mathrm{Var}(Y_{t+1}|D_t) \\
&= -aW_2 + aW_2 + \mathrm{Var}(Y_{t+1}|D_t) \\
&= \mathrm{Var}(Y_{t+1}|D_t)\,.
\end{aligned}$$

In general, from equation 3.6 we have that

$$\begin{aligned}\mathrm{Var}(Y'_{t+k}|D_t) &= \mathrm{Var}(Y_{t+k}|D_t) + \mathbf{E}_2^T\mathbf{J}_2(1)^k\delta\mathbf{C}(\mathbf{J}_2(1)^T)^k\mathbf{E}_2 \\ &+ \mathbf{E}_2^T\mathbf{J}_2(1)^{k-1}\delta\mathbf{W}(\mathbf{J}_2(1)^T)^{k-1}\mathbf{E}_2 + \ldots + \mathbf{E}_2^T\delta\mathbf{W}\mathbf{E}_2 \ ;\end{aligned}$$

with $\mathbf{J}_2(1)^k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$, this gives

$$\begin{aligned}\mathbf{E}_2^T\mathbf{J}_2(1)^k\delta\mathbf{C}(\mathbf{J}_2(1)^T)^k\mathbf{E}_2 &= -k^2aW_2 \\ \text{and } \mathbf{E}_2^T\mathbf{J}_2(1)^j\delta\mathbf{W}(\mathbf{J}_2(1)^T)^j\mathbf{E}_2 &= (2j+1)aW_2 \ .\end{aligned}$$

Hence

$$\begin{aligned}\sum_{j=0}^{k-1}\mathbf{E}_2^T\mathbf{J}_2(1)^j\delta\mathbf{W}(\mathbf{J}_2(1)^T)^j\mathbf{E}_2 &= k^2aW_2 \\ \Rightarrow \mathrm{Var}(Y'_{t+k}|D_t) &= \mathrm{Var}(Y_{t+k}|D_t) \ .\end{aligned}$$

It is equally easy to show that $\mathrm{Cov}(Y'_{t+j}, Y'_{t+k}|D_t) = \mathrm{Cov}(Y_{t+j}, Y_{t+k}|D_t)$ from equation 3.5, and this tells us that under M and M' the distribution of the forecast function $f_t(k)$ is identical for all time points.

Given our initial intention of solving for $\mathbf{W}$ uniquely, this is something of an unexpected result. It shows that there are an infinite number of possible choices for $\mathbf{W}$ which will all give *identical* long-term or converged forecast distributions - so trying to solve uniquely for $\mathbf{W}$ is, quite simply, infeasible. Whatever approach we adopt, we will always be left with one degree of freedom unfixed in $\mathbf{W}$. There is a vital interpretation of this example too: whatever $\mathbf{W}'$ we should decide to

choose from the canonical class, we can write as

$$\mathbf{W}' = \begin{pmatrix} W_1 + aW_2 & aW_2 \\ aW_2 & W_2 \end{pmatrix} = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix} + \begin{pmatrix} aW_2 & aW_2 \\ aW_2 & 0 \end{pmatrix}$$

$$= \mathbf{W} + \delta\mathbf{W}$$

i.e. we can fully define this long-term equivalent (as in the definition of section 2.10) family of models by solving for $\mathbf{W}$, which is *diagonal* (presuming its variance matrix form is still satisfied). It is only in West and Harrison [36] that the hesitant proposition is made, that these ideas "seem to suggest that any specified model can be transformed to one with a particularly simple form, based on a *diagonal* evolution matrix". As they state, this can be done in many cases, but they are then distracted by considering the model $\{(1,1), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, V, \mathbf{W}\}$, for which we can always choose

$$\mathbf{W}' = \begin{pmatrix} W_1 & W_3 \\ W_3 & W_2 \end{pmatrix} + \begin{pmatrix} a & -a \\ -a & a \end{pmatrix}$$

$$= \mathbf{W} + \delta\mathbf{W}$$

to give identical forecast distributions. Here we see that $\mathbf{W}$ is not diagonal - however, note that the TSDLM is *not* observable, since $\mathbf{G} = \mathbf{I}_2$, and hence $\mathbf{T} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ of rank 1, as we commented is always the case for $\mathbf{G} = \mathbf{I}_p$ in section 2.8. Whilst we find that it is still not always possible to define a diagonal $\mathbf{W}$ in all constant TSDLMs, by considering just *observable* models we shall now prove the valuable result that it is never feasible to solve uniquely for more than p elements of $\delta\mathbf{W}$, and hence that it is only ever possible to deter-

mine p degrees of freedom in $\mathbf{W}$. Therefore, so long as its variance matrix form is still satisfied, we can reduce $\mathbf{W}$ to a diagonal form in any observable TSDLM.

Lemma 3.1: We define two observable, constant TSDLMS, which have the same canonical form and which, further, exhibit long-term equivalence as defined in section 2.10. These two models are given by

$$\begin{aligned} M &= \{\mathbf{E}, \mathbf{J}, V, \mathbf{W}\} \\ \text{and } M' &= \{\mathbf{E}, \mathbf{J}, V, \mathbf{W}' = \mathbf{W} + \delta\mathbf{W}\}\ , \\ \text{for } \mathbf{E} &= (\mathbf{E}_{r_1}, \ldots, \mathbf{E}_{r_s})^T \\ \mathbf{E}_p &= (1, 0, \ldots, 0)^T\ , \quad \text{a p} - \text{vector}, \\ \text{and } \mathbf{J} &= \text{blockdiag}[\mathbf{J}_{r_1}(\lambda_1), \ldots, \mathbf{J}_{r_s}(\lambda_s)] \end{aligned}$$

with $\mathbf{J}_p(\lambda)$ a $p \times p$ Jordan block, all as defined in section 2.9. From results in chapter 2 we know that under M, $\mathbf{C}_t \to \mathbf{C}$; $\mathbf{A}_t \to \mathbf{A}$; $\mathbf{R}_t \to \mathbf{R}$ and $Q_t \to Q$, and hence under M', $\mathbf{C}_t \to \mathbf{C} + \delta\mathbf{C}$; $\mathbf{A}_t \to \mathbf{A} + \delta\mathbf{A}$; $\mathbf{R}_t \to \mathbf{R} + \delta\mathbf{R}$ and $Q_t \to Q + \delta Q$. Then the perturbations in $\mathbf{W}$ and $\mathbf{C}$, namely $\delta\mathbf{W}$ and $\delta\mathbf{C}$, satisfy

$$\delta\mathbf{W} = \delta\mathbf{C} - \mathbf{J}\delta\mathbf{C}\mathbf{J}^T\ .$$

Proof: We consider the k^{th} step-ahead forecast equations again. Since the models are long-term equivalent, these distributions are identical under both M and M', and so we have that

$$\text{Var}(Y_{t+k}|D_t, M) = \text{Var}(Y'_{t+k}|D_t, M')$$

$$\Rightarrow \mathbf{E}^T(\mathbf{J}^k\mathbf{C}(\mathbf{J}^T)^k+\mathbf{J}^{k-1}\mathbf{W}(\mathbf{J}^T)^{k-1}+\ldots+\mathbf{J}\mathbf{W}\mathbf{J}^T+\mathbf{W})\mathbf{E} = \mathbf{E}^T(\mathbf{J}^k(\mathbf{C}+\delta\mathbf{C})(\mathbf{J}^T)^k$$
$$+\mathbf{J}^{k-1}(\mathbf{W}+\delta\mathbf{W})(\mathbf{J}^T)^{k-1}+\ldots+(\mathbf{W}+\delta\mathbf{W}))\mathbf{E}$$

$$\Rightarrow \mathbf{E}^T(\mathbf{J}^k\delta\mathbf{C}(\mathbf{J}^T)^k + \mathbf{J}^{k-1}\delta\mathbf{W}(\mathbf{J}^T)^{k-1}+\ldots+\delta\mathbf{W})\mathbf{E} = 0\ .$$

But we must also have

$$\mathrm{Var}(Y_{t+k-1}|D_t, M) = \mathrm{Var}(Y'_{t+k-1}|D_t, M')$$

$$\Rightarrow \mathbf{E}^T(\mathbf{J}^{k-1}\delta\mathbf{C}(\mathbf{J}^T)^{k-1} + \mathbf{J}^{k-2}\delta\mathbf{W}(\mathbf{J}^T)^{k-2}+\ldots+\delta\mathbf{W})\mathbf{E} = 0\ ,$$

and so subtracting these two equations gives

$$\begin{aligned}\mathbf{E}^T(\mathbf{J}^k\delta\mathbf{C}(\mathbf{J}^T)^k + \mathbf{J}^{k-1}(\delta\mathbf{W}-\delta\mathbf{C})(\mathbf{J}^T)^{k-1})\mathbf{E} &= 0\\ \Rightarrow \mathbf{E}^T\mathbf{J}^{k-1}(\mathbf{J}\delta\mathbf{C}\mathbf{J}^T + (\delta\mathbf{W}-\delta\mathbf{C}))(\mathbf{J}^T)^{k-1}\mathbf{E} &= 0\ . \qquad (3.7)\end{aligned}$$

By looking at the covariances of Y'_{t+j}, Y'_{t+k} from 3.5 we come to a similar equation, namely

$$\mathbf{E}^T\mathbf{J}^k(\mathbf{J}\delta\mathbf{C}\mathbf{J}^T + (\delta\mathbf{W}-\delta\mathbf{C}))(\mathbf{J}^T)^j\mathbf{E} = 0\ .$$

If we consider first the simpler model definition of $\mathbf{J} \equiv \mathbf{J}_p(\lambda)$ (one eigenvalue λ, multiplicity p), so that $\mathbf{E} \equiv \mathbf{E}_p = (1, 0, \ldots, 0)^T$, we find that

$$\mathbf{J}^k \equiv (\mathbf{J}_p(\lambda))^k = \left(J^k_{ij}\right)_{p\times p}\ ,$$

where

$$\left.\begin{array}{ll} J_{ij}^k = 0 & (i > j) \\ J_{ii}^k = \lambda^k & (i = j) \\ J_{ij}^k = \binom{k}{j-i} \lambda^{k-(j-i)} & (0 < (j-i) \leq k) \\ J_{ij}^k = 0 & ((j-i) > k) \end{array}\right\} \text{ for } k \leq p-1 \, .$$

(i.e. $(\mathbf{J}_p(\lambda))^k$ has $k+1$ non-zero diagonals of constants, starting on the main diagonal of λ^k, the superdiagonal of $k\lambda^{k-1}$, etc.). Hence

$$\mathbf{E}^T\mathbf{J}^k = \mathbf{E}_p^T(\mathbf{J}_p(\lambda))^k = (\lambda^k, k\lambda^{k-1}, \ldots, \binom{k}{j}\lambda^{k-j}, \ldots, k\lambda, 1, 0, \ldots, 0) \, , \; k \leq p-1,$$

is a p-vector.

This leads us to the conclusion that $\mathbf{E}_p^T(\mathbf{J}_p(\lambda))^k$, for $k = 0, \ldots, p-1$, are all linearly independent (and so, trivially, are $(\mathbf{J}_p(\lambda)^T)^k\mathbf{E}_p$). Therefore equation 3.7 can hold for $k = 1, \ldots, p$ iff

$$\mathbf{J}_p(\lambda)\delta\mathbf{C}(\mathbf{J}_p(\lambda))^T + (\delta\mathbf{W} - \delta\mathbf{C}) = \mathbf{0} \, .$$

So, returning to the definitions of M and M', by recalling that the general Jordan form is $\mathbf{J} = \text{blockdiag}[\mathbf{J}_{r_1}(\lambda_1), \ldots, \mathbf{J}_{r_s}(\lambda_s)]$ and $\mathbf{E} = (\mathbf{E}_{r_1}, \ldots, \mathbf{E}_{r_s})^T$ we see that this result will hold for each block element of $\mathbf{J}$ in turn, due to the partitioning of $\mathbf{J}$ and $\mathbf{E}$. Hence we must have that 3.7 holds for $k = 1, \ldots, p$ if and only if

$$\mathbf{J}\delta\mathbf{C}\mathbf{J}^T + (\delta\mathbf{W} - \delta\mathbf{C}) = \mathbf{0} \, ,$$

giving the required result. •

Lemma 3.1 has achieved two things. Firstly, since $\delta\mathbf{R} = \mathbf{J}\delta\mathbf{C}\mathbf{J}^T + \delta\mathbf{W}$, it shows that $\delta\mathbf{R} = \delta\mathbf{C}$ in our long-term equivalent model system. Secondly, it allows us to calculate the additional $\delta\mathbf{W}$ in $\mathbf{W}'$ given a perturbation of $\delta\mathbf{C}$ to $\mathbf{C}$ in M'. Note that the converse calculation is not applicable, since we cannot compute $\delta\mathbf{C}$ from $\delta\mathbf{W}$, given the form of Lemma 3.1 - instead, we must use Lemma 3.2.

Lemma 3.2: With the two long-term equivalent models as before, M and M', as long as the models are non-degenerate (so that $\mathbf{RE} \neq \mathbf{0}$ and we cannot express certain elements of $\boldsymbol{\theta}_t$ in terms of others) the perturbation in the convergent form of $\mathbf{C}'$ in M', $\delta\mathbf{C}$, satisfies the relation

$$\mathbf{E}^T\delta\mathbf{C} = \mathbf{0} .$$

Proof: From the Kalman Filter recurrence equations 2.3, we have

$$\mathbf{C}' = \mathbf{R}' - \frac{\mathbf{R}'\mathbf{E}\mathbf{E}^T\mathbf{R}'}{Q'} \tag{3.8}$$

$$\begin{aligned} \Rightarrow \mathbf{C} + \delta\mathbf{C} &= \mathbf{R} + \delta\mathbf{R} - \frac{(\mathbf{R} + \delta\mathbf{R})\mathbf{E}\mathbf{E}^T(\mathbf{R} + \delta\mathbf{R})}{Q + \delta Q} \\ &= \mathbf{R} + \delta\mathbf{R} - \frac{\mathbf{R}\mathbf{E}\mathbf{E}^T\mathbf{R}}{Q + \delta Q} - \frac{\delta\mathbf{R}\mathbf{E}\mathbf{E}^T\mathbf{R} + \mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R} + \delta\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R}}{Q + \delta Q} . \end{aligned}$$

However, as M and M' are long-term equivalent, $\mathrm{Var}(Y'_{t+1}|D_t, M') = Q' =$

$Q + \delta Q = \mathrm{Var}(Y_{t+1}|D_t, M)$, so that $\delta Q = 0$ and thus

$$\begin{aligned}
\mathbf{C} + \delta\mathbf{C} &= \mathbf{R} - \frac{\mathbf{R}\mathbf{E}\mathbf{E}^T\mathbf{R}}{Q} + \delta\mathbf{R} - \frac{\delta\mathbf{R}\mathbf{E}\mathbf{E}^T\mathbf{R} + \mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R} + \delta\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R}}{Q} \\
\Rightarrow \delta\mathbf{C} &= \delta\mathbf{R} - \frac{\delta\mathbf{R}\mathbf{E}\mathbf{E}^T\mathbf{R} + \mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R} + \delta\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R}}{Q} \\
&= \delta\mathbf{R} - \frac{2\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R} + \delta\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R}}{Q} \quad \text{(since } \mathbf{R} \text{ and } \delta\mathbf{R} \text{ are symmetric)} \\
\Rightarrow Q\delta\mathbf{C} &= Q\delta\mathbf{R} - 2\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R} - \delta\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{R} \,.
\end{aligned}$$

But from Lemma 3.1, $\delta\mathbf{R} = \delta\mathbf{C}$, and so

$$2\mathbf{R}\mathbf{E}\mathbf{E}^T\delta\mathbf{C} + \delta\mathbf{C}\mathbf{E}\mathbf{E}^T\delta\mathbf{C} = \mathbf{0} \;;$$

since $2\mathbf{R}\mathbf{E}$ is not necessarily $\mathbf{0}$ (and, evidently, not necessarily equal to $-\delta\mathbf{C}\mathbf{E}$ either), we require $\mathbf{E}^T\delta\mathbf{C} = \mathbf{0}$, the desired result. •

This leads us to the following theorem.

Theorem 3.3: With the two long-term equivalent models $M = \{\mathbf{E}, \mathbf{J}, V, \mathbf{W}\}$ and $M' = \{\mathbf{E}, \mathbf{J}, V, \mathbf{W}' = \mathbf{W} + \delta\mathbf{W}\}$, we can only fix at most p degrees of freedom in $\mathbf{W}$. Hence, as long as its variance matrix form is satisfied, we can express the evolution variance matrix $\mathbf{W}$, which defines the long-term equivalence class, in a *diagonal* form.

The proof follows easily from Lemmas 3.1 and 3.2. $\mathbf{E}^T\delta\mathbf{C} = \mathbf{0}$, together

with $\delta\mathbf{C} = \delta\mathbf{R}$ from Lemma 3.1, gives us that

$$\delta Q = \mathbf{E}^T \delta\mathbf{R}\mathbf{E} = 0 \ ,$$

hence enabling us to reverse the proof of Lemma 3.2 and derive the Kalman Filter form 3.8 for $\mathbf{C}'$ from $\mathbf{E}^T\delta\mathbf{C} = \mathbf{0}$. So Lemma 3.2 produces a set of p linearly independent equations in the elements of $\delta\mathbf{C}$ (where $\boldsymbol{\theta}_t$, and hence $\mathbf{E}$, is p-dimensional), and it is indeed impossible to determine more than these p degrees of freedom in $\delta\mathbf{C}$. Further, from Lemma 3.1 and $\delta\mathbf{W} = \delta\mathbf{C} - \mathbf{J}\delta\mathbf{C}\mathbf{J}^T$, we are only able to fix exactly p dimensions in $\delta\mathbf{W}$, and similarly in $\mathbf{W}'$. Hence we will always be able to write $\mathbf{W}'$ as $\mathbf{W} + \delta\mathbf{W}$ for $\mathbf{W}$ having exactly p non-zero elements, which we can choose to lie entirely on the diagonal as long as the variance matrix requirements on $\mathbf{W}$ are still satisfied. •

The potentially diagonal $\mathbf{W}$ defines the long-term equivalence class of models which all have a particular forecast function; the matrix $\delta\mathbf{W}$, which is calculable from Lemmas 3.1 and 3.2, then spans this entire class with $\mathbf{W}$.

Example. Again, we consider the canonical model

$$M = \left\{(1,0), \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, V, \mathbf{W}\right\} .$$

Then Lemma 3.2 implies

$$(1,0)\begin{pmatrix} \delta C_1 & \delta C_3 \\ \delta C_3 & \delta C_2 \end{pmatrix} = \mathbf{0}$$

$$\Rightarrow \delta C_1 = 0 = \delta C_3$$

$$\text{and so } \delta\mathbf{C} = \begin{pmatrix} 0 & 0 \\ 0 & \delta C_2 \end{pmatrix},$$

the form we took for $\delta\mathbf{C}$ in the earlier citation of this model. Next, $\delta\mathbf{W}$ is found from Lemma 3.1:

$$\begin{aligned} \delta\mathbf{W} = \delta\mathbf{C} - \mathbf{J}\delta\mathbf{C}\mathbf{J}^T &= \delta\mathbf{C} - \begin{pmatrix} \delta C_2 & \delta C_2 \\ \delta C_2 & \delta C_2 \end{pmatrix} \\ &= \begin{pmatrix} -\delta C_2 & -\delta C_2 \\ -\delta C_2 & 0 \end{pmatrix}, \end{aligned}$$

which is again in the form given earlier. So we can write any evolution matrix $\mathbf{W}'$ in this long-term equivalence class as $\mathbf{W}+\delta\mathbf{W}$, for $\mathbf{W}$ equal to $\begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}$ (which is a valid diagonal form for this variance matrix so long as W_1 and W_2 are both ≥ 0), and further, for $\delta\mathbf{W}$ of the form $\begin{pmatrix} aW_2 & aW_2 \\ aW_2 & 0 \end{pmatrix}$; the particular class is uniquely determined - in other words, the forecast distribution is itself determined - by the choice of the (potentially diagonal) evolution variance matrix $\mathbf{W}$. (Note that this diagonalisation is always feasible when, for $\mathbf{W}' = \begin{pmatrix} W_1' & W_3' \\ W_3' & W_2' \end{pmatrix}$, we have $W_3' \leq W_1'$, for then $\delta\mathbf{W} = \begin{pmatrix} W_3' & W_3' \\ W_3' & 0 \end{pmatrix}$ and so $\mathbf{W} = \begin{pmatrix} W_1 = W_1' - W_3' & 0 \\ 0 & W_2 = W_2' \end{pmatrix}$, which is a valid variance matrix when $W_1 = W_1' - W_3' \geq 0$.)

3.7 Comparison of the discounting approach with specification of $\mathbf{W}_t$

When assessing whether or not to model sequential loss of information about our state vector by employing the use of a discounting matrix $\mathbf{B}_t$, we must consider the following two points:

(i) are we likely to have extreme levels of knowledge of the state, $\boldsymbol{\theta}_t$, at any stage of the analysis? Having either very precise knowledge or, contrastingly, very vague information about $\boldsymbol{\theta}_t$ will lead to, in both instances, a fixed percentage discount of this information being an unrealistic absolute value to include in $\mathbf{R}_t = \mathbf{B}_t^{-1/2}\mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T\mathbf{B}_t^{-1/2}$.

(ii) Can we assign realistic values to the diagonal matrix $\mathbf{B}_t$ that will cover the (almost certainly) differing nature of the elements of $\boldsymbol{\theta}_t$?

If the answer to (i) is "Yes", then the concepts of the discounting approach are not only a poor approximation for the actual processes involved in sequential information loss, but are also more than likely to lead us into major practitional difficulties. And even if the answer to this question is "No", the conscientious practitioner will not be happy with just choosing $\delta_i = 0.95$-or-so throughout his discounting matrix $\mathbf{B}_t$; when faced with question (ii) he is likely to start feeling uncomfortable... "Are all my discount parameters the same? If so, why? Surely parameter X has more sequential variation than parameter Y? And is 3% or 5% a better representation of my loss in initial information? For that matter, what *is* 3% of my initial information, and what will it be at time t? Hmmmm, what was question (i) again?..."

Once we start analysing question (ii) at length, it soon starts to beg all the further questions that, with a different interpretation, are required to be answered before we can develop an additive loss of information, in taking $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{W}_t$. And since the additive $\mathbf{W}_t$ approach to the problem automatically sidesteps question (i), we are naturally led into discarding the possibility of discounting when this issue is of serious contemplation.

The only *conceptual* criticism of attempting to specify a $\mathbf{W}_t$ matrix lies in its non-uniqueness that we have just illustrated. However, now that we have shown that we can simplify the definition of any observable, constant TSDLM to one involving a potentially diagonal $\mathbf{W}$, this non-uniqueness has, in fact, become an advantage in consideration of these models - in appropriate models we need only consider the specification of the easier-to-interpret variances within $\mathbf{W}$, and can ignore the covariances.

In their reply to discussions arising from their paper, Ameen and Harrison stress that the Bayesian approach to modelling allows us the freedom to express "the way in which we wish to view the data", and that it is "totally undesirable to be imprisoned by such concepts as stationarity". The undeniable truth in these comments is certainly not in question. The point of this discussion is to highlight an area of the Bayesian approach which is undoubtedly a major difficulty, and to endeavour to ensure that it is not swept under the carpet or left in the hands of a black-box method, just because it is conceptually simple. These are the exact types of black-box methods that the Bayesian should be endeavouring to free himself from! But for Jack-the-practitioner, these conceptual arguments and interpretations are of secondary importance to

the pressing question of "which approach do I choose for now?". And it is not doubted that there will be many instances where he will conclude, with care, that it is desirable to represent the loss of information from one time point to the next by a fixed percentage of that information, and hence use a discounting matrix. But if he simply chooses to follow this method to side-step the problem he is immediately imprisoned yet again, only this time by a concept which has been illustrated, in a practical case, to be potentially wholly inappropriate; with a little more thought the loss of information can be modelled in a far more effective and *realistic* manner.

But if the practitioner should decide that the answer to the first question raised above is, after all, "Yes", and hence that discounting is not appropriate, by opting to be more certain and thus trying to specify an additive $\mathbf{W}_t$ he has not yet solved his problem in entirety. In the constant TSDLMs which so many practitioners do, and always will, use, he must still specify some p diagonal elements, and even this is a far from simple task, however he should decide to tackle it (specifying the diagonal elements of $\mathbf{W}_t$ in the earlier advertising awareness example actually took much careful consideration, which of course is not apparent from the description of the analysis).

And so we return to the earlier method of looking at step-ahead forecast variances and covariances to estimate V and $\mathbf{W}$. To avoid the major pitfall of divergence due to sub-optimal filtering, in the next chapter we make several calculations in the constant TSDLM relating to the exact effects of model mis-specification on this forecast distribution, enabling 'our Jack' to estimate the true values of V and $\mathbf{W}$ *whatever* initial specifications he should first choose.

Chapter 4

Estimation of V and $\mathbf{W}$ in the Constant TSDLM

4.1 Divergence of estimates due to sub-optimal filtering

In the many examples cited in section 3.5 that attempt to estimate V and $\mathbf{W}$ through the forecast distribution equations 3.5 and 3.6, there is a common theme to be found throughout. They all run the not-inconsiderable risk of divergence of these on-line variance estimates due to their use of sub-optimal filtering techniques.

We have already discussed, at the end of section 3.1, how the performance of the DLM is extremely sensitive to the choices made of $\mathbf{W}_t$ and V_t. And as we shall see further in section 4.3.3, the effect of underestimating the signal-to-noise ratio, $r = W/V$, for instance (i.e. taking W too small with respect to V), in the

steady model, leads to the variance of the k^{th} step-ahead forecast errors tending to infinity, along with $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t)$. Hence specifying too small a W leads to the LHS of equation 3.5 inflating rapidly, from whence our feedback estimate $\widehat{W}$ of W, calculated *directly* from 3.5, can be far too large, even assuming that we have correctly specified the observational variance V_t originally. This correct specification is, of course, extremely unlikely, and we can be led into an even more meaningless situation very easily: by overestimating V, we increase the limiting value of $C = AV$ (see section 4.3.2), which when subtracted from the estimate of $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t)$ can give a large *negative* value for $\widehat{W}$.

In the more general DLM specification, underspecifying elements in $\mathbf{W}_t$ with respect to V_t will mean that the data is largely 'ignored' during the analysis, since any perturbations in the data are accommodated by the model through the relatively large value of V_t, and not an underlying shift in the state vector $\boldsymbol{\theta}_t$. This particular misspecification of $\mathbf{W}_t$, therefore, results in little change in the posterior $\mathbf{m}_{t+1}$ with respect to $\mathbf{G}_{t+1}\mathbf{m}_t$ (qualitatively due to $\mathbf{R}_t$ being small relative to Q_t, and so the adaptive coefficient $\mathbf{A}_t$ is small). Thus the variance of the step-ahead forecast function will inflate rapidly under this $\{\mathbf{W}_t, V_t\}$ specification, along with the LHS of equation 3.5. This makes the evolution of the on-line feedback estimates $\{\widehat{\mathbf{W}}_t\}$, at best, extremely slow in its convergence; it is more likely, in fact, that we will be faced with divergence of these estimates. Even following the preferred course of overestimating the elements of $\mathbf{W}_t$, so that the model 'over-adapts' (due to the large adaptive coefficient $\mathbf{A}_t$) to perceived changes in the data evolution (resulting in $\mathbf{F}_{t+1}^T\mathbf{m}_{t+1}$ relatively closer to Y_{t+1} than $\mathbf{F}_{t+1}^T\mathbf{G}_{t+1}\mathbf{m}_t = f_{t+1}$), does not guarantee us sensible convergence of

$\widehat{\mathbf{W}}$, as we shall see again in section 4.3 when studying the steady model: whilst keeping the step-ahead forecast error variance function finite, we still introduce a (potentially large) bias into the LHS of 3.5.

We find that the well-behaved limiting properties of the constant TSDLM, $\{\mathbf{F}, \mathbf{G}, V, \mathbf{W}\}$, allow us to produce equations that are *exactly* soluble algebraically for V and $\mathbf{W}$, *irrespective of our initial model definition.*

4.2 Effects of model misspecification

Suppose that the data evolves through a particular constant TSDLM

$$M_0 = \{\mathbf{F}, \mathbf{G}, V_0, \mathbf{W}_0\} ,$$

with true variances V_0 and $\mathbf{W}_0$; it is then our aim to estimate these true values. We will presume that the practitioner has correctly identified the evolution of the data - i.e. he has defined a model with correct[1] $\mathbf{F}$ and $\mathbf{G}$ - and has made initial specifications for the observation and state evolution variances of V and $\mathbf{W}$ in a model M. Further, we reiterate that the model is presupposed to have reached its *convergent* form. Then recall from section 2.8 that this general constant TSDLM of

$$M = \{\mathbf{F}, \mathbf{G}, V, \mathbf{W}\}$$

[1]This is, of course, not usually possible, as we are only expressing our views (normally quite simplistic for parsimony) on how we would wish to model the data evolution; see, for instance, Harrison and Stevens [18], or West and Harrison [36] amongst others for more details on this discussion.

can be written in its limiting form as equation 2.11

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + e_t + \sum_{j=1}^{p} \beta_j e_{t-j} \ ,$$

where

$$\alpha_1 = \sum_{i=1}^{p} \lambda_i \ ; \quad \alpha_2 = -\sum_{i=1}^{p} \sum_{k=i+1}^{p} \lambda_i \lambda_k \ ; \ \ldots \ ; \quad \alpha_p = (-1)^p \lambda_1 \ldots \lambda_p$$

$$\text{and} \quad \beta_1 = -\sum_{i=1}^{p} \rho_i \ ; \quad \beta_2 = \sum_{i=1}^{p} \sum_{k=i+1}^{p} \rho_i \rho_k \ ; \ \ldots \ ; \quad \beta_p = (-1)^{p+1} \rho_1 \ldots \rho_p \ ,$$

for λ_i, ρ_i the eigenvalues of $\mathbf{G}$ and $\mathbf{H} = (\mathbf{I} - \mathbf{A}\mathbf{F}^T)\mathbf{G} = \mathbf{C}\mathbf{R}^{-1}\mathbf{G}$ respectively. From this model form we have

$$\begin{aligned}
\mathrm{E}[Y_{t+1}|D_t] &= \alpha_1 Y_t + \ldots + \alpha_p Y_{t+1-p} + \beta_1 e_t + \ldots + \beta_p e_{t+1-p} \ , \\
\mathrm{E}[Y_{t+2}|D_t] &= \alpha_1 \mathrm{E}[Y_{t+1}|D_t] + \sum_{j=2}^{p} \alpha_j Y_{t+2-j} + \sum_{j=2}^{p} \beta_j e_{t+2-j} \\
\text{and so on, to} \quad \mathrm{E}[Y_{t+p}|D_t] &= \sum_{j=1}^{p} \mathrm{E}[Y_{t+p-j}|D_t] + \alpha_p Y_t + \beta_p e_t \ .
\end{aligned}$$

We now introduce some notation. Henceforth, we consider these step-ahead forecasts under the two different models M_0 (the true model) and M separately, and so take the expectations conditioned not only upon D_t, but the relevant variance specifications $\{V, \mathbf{W}\}$ or $\{V_0, \mathbf{W}_0\}$. We also require a distinction between our symbolic notations under M and M_0; where a subscript is not already present on a symbol, we will simply make this distinction via the presence or otherwise of a subscript '0', and where a subscript is already present, we shall use a circumflex ('hat') on all symbols pertaining to the estimated model M

(whilst leaving those under M_0 as they are). Further, we adapt some Box-Jenkins notation slightly by denoting the k^{th} step-ahead forecast error at time t, $Y_{t+k} - \mathrm{E}[Y_{t+k}|D_t]$, by $\hat{e}_t(k|D_t, V, \mathbf{W})$ under M (and so by $e_t(k|D_t, V_0, \mathbf{W}_0)$ under M_0); this notation is then abbreviated additionally (and *only*) for the *one* step-ahead forecast error at time t, where we denote $e_t(1|D_t, V_0, \mathbf{W}_0) \equiv e_{t+1}$ and $\hat{e}_t(1|D_t, V, \mathbf{W}) \equiv \hat{e}_{t+1}$, so that $e_t \equiv a_t$ and $\hat{e}_t \equiv e_t$ in relation to the Box-Jenkins notation.

Thus proceeding to define

$$\sum_{j=i}^{p}(\hat{\alpha}_j Y_{t+i-j} + \hat{\beta}_j \hat{e}_{t+i-j}) \equiv \widehat{\xi}_i$$

under M (as opposed to $\sum_{j=i}^{p}(\alpha_j Y_{t+i-j} + \beta_j e_{t+i-j}) \equiv \xi_i$ under M_0), we have

$$\begin{aligned}
\mathrm{E}[Y_{t+1}|D_t, V, \mathbf{W}] &= \widehat{\xi}_1 \,, \\
\mathrm{E}[Y_{t+2}|D_t, V, \mathbf{W}] &= \hat{\alpha}_1 \mathrm{E}[Y_{t+1}|D_t, V, \mathbf{W}] + \widehat{\xi}_2 \\
&= \hat{\alpha}_1 \widehat{\xi}_1 + \widehat{\xi}_2 \,, \\
\mathrm{E}[Y_{t+3}|D_t, V, \mathbf{W}] &= \hat{\alpha}_1 \mathrm{E}[Y_{t+2}|D_t, V, \mathbf{W}] + \hat{\alpha}_2 \mathrm{E}[Y_{t+1}|D_t, V, \mathbf{W}] + \widehat{\xi}_3 \\
&= (\hat{\alpha}_1^2 + \hat{\alpha}_2)\widehat{\xi}_1 + \hat{\alpha}_1 \widehat{\xi}_2 + \widehat{\xi}_3 \,, \\
&\vdots \\
\Rightarrow \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}] &= \sum_{i=1}^{k} \Lambda_{k-i} \widehat{\xi}_i \,, \quad k = 1, \ldots, p, \qquad (4.1)
\end{aligned}$$

where

$$\Lambda_k = \sum_{j=1}^{k} \hat{\alpha}_j \Lambda_{k-j} \,, \quad \text{for } \Lambda_0 = 1. \qquad (4.2)$$

These observations lead us to the following Lemma:

Lemma 4.1: The difference in the expectation of Y_{t+k} under the two models M and M_0 is given by

$$\mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}] = \sum_{i=1}^{k} \Lambda_{k-i} \left(\sum_{j=i}^{p} \left(\beta_j - \hat{\beta}_j \left(\frac{1 + \sum_{l=1}^{p} \beta_l \mathrm{B}^l}{1 + \sum_{l=1}^{p} \hat{\beta}_l \mathrm{B}^l} \right) \right) e_{t+i-j} \right), \tag{4.3}$$

for $1 \leq k \leq p$, and where B is the backward shift operator (so that $Be_t = e_{t-1}$, etc.; note also that the polynomial ratio $\left(\frac{1+\sum_{l=1}^{p} \beta_l \mathrm{B}^l}{1+\sum_{l=1}^{p} \hat{\beta}_l \mathrm{B}^l} \right)$ acts in single combination on e_{t+i-j}).

Proof: We remark that the eigenvalues λ_j of $\mathbf{G}$ under M_0 are the same as those under M, since we are presuming the practitioner has specified the correct system evolution matrix, whereas the eigenvalues of $\mathbf{H} = (\mathbf{I} - \mathbf{A}\mathbf{F}^T)\mathbf{G}$ and $\mathbf{H}_0 = (\mathbf{I} - \mathbf{A}_0\mathbf{F}^T)\mathbf{G}$ are different, as the convergent values of $\mathbf{A}$ and $\mathbf{A}_0$ under M and M_0 are influenced by the choice of variances in the model. Thus $\hat{\alpha}_j = \alpha_j$ for all j and so the Λ_j are identical under either model (since they are both functions of the λ_i only), whereas $\hat{\beta}_j$ and β_j are distinct under M and M_0; from equation 4.1 this produces

$$\begin{aligned}
&\mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}] = \sum_{i=1}^{k} \Lambda_{k-i} \left(\xi_i - \hat{\xi}_i \right) \\
= &\sum_{i=1}^{k} \Lambda_{k-i} \left(\sum_{j=i}^{p} (\alpha_j Y_{t+i-j} + \beta_j e_{t+i-j}) - \sum_{j=i}^{p} \left(\alpha_j Y_{t+i-j} + \hat{\beta}_j \hat{e}_{t+i-j} \right) \right) \\
= &\sum_{i=1}^{k} \Lambda_{k-i} \left(\sum_{j=i}^{p} \left(\beta_j e_{t+i-j} - \hat{\beta}_j \hat{e}_{t+i-j} \right) \right) .
\end{aligned}$$

Now, in their limiting forms we can equate the two models, and so from 2.11

we have

$$Y_t - \sum_{j=1}^{p} \alpha_j Y_{t-j} = \hat{e}_t + \sum_{j=i}^{p} \hat{\beta}_j \hat{e}_{t-j} \;=\; e_t + \sum_{j=1}^{p} \beta_j e_{t-j}$$

$$\Rightarrow \left(1 + \sum_{j=1}^{p} \hat{\beta}_j \mathrm{B}^j\right) \hat{e}_{t+1} \;=\; \left(1 + \sum_{j=1}^{p} \beta_j \mathrm{B}^j\right) e_{t+1} \; .$$

Substitution for $\hat{e}_{t+i-j}$ in the above equation from this limiting form then completes the proof.•

Now notice that

$$\begin{aligned} Y_{t+k} - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}] \;&=\; Y_{t+k} - \mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] \\ &+\; \mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}] \end{aligned}$$

$$\Rightarrow \hat{e}_t(k|D_t, V, \mathbf{W}) = e_t(k|D_t, V_0, \mathbf{W}_0) + \mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}] \; . \tag{4.4}$$

In the proof of Lemma 4.1 we were only concerned with the latter part of this equation, $\mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}]$. Now we can return to the other half of equation 4.4. We would like to find $e_t(k|D_t, V_0, \mathbf{W}_0)$ in terms of e_{t+i}'s that are independent of 4.3, since this will ease our later calculations of $\mathrm{Var}(\hat{e}_t(k|D_t, V, \mathbf{W}))$ and $\mathrm{Cov}(Y_{t+j}, Y_{t+k}|D_t, V, \mathbf{W})$ considerably. This wish leads to the next Lemma.

Lemma 4.2: Under the defining model M_0,

$$e_t(k|D_t, V_0, \mathbf{W}_0) = -\sum_{i=1}^{k} \left(\left(\sum_{j=0}^{k-i} \Lambda_j \beta_{k-i-j} \right) e_{t+i} \right) , \tag{4.5}$$

for $\beta_0 = -1$ and $1 \leq k \leq p$.

Proof: Equation 4.1 produces

$$\begin{aligned} e_t(k|D_t, V_0, \mathbf{W}_0) &= Y_{t+k} - \sum_{i=1}^{k} \Lambda_{k-i} \xi_i \\ \text{where } \xi_i &= Y_{t+i} - e_{t+i} - \sum_{j=1}^{i-1} (\alpha_{i-j} Y_{t+j} - \beta_{i-j} e_{t+j}) , \end{aligned}$$

and so

$$\begin{aligned} e_t(k|D_t, V_0, \mathbf{W}_0) &= Y_{t+k} - \sum_{i=1}^{k} \Lambda_{k-i} \left(Y_{t+i} - e_{t+i} - \sum_{j=1}^{i-1} (\alpha_{i-j} Y_{t+j} - \beta_{i-j} e_{t+j}) \right) \\ &= e_{t+k} - \sum_{i=1}^{k-1} \left(\begin{array}{r} \Lambda_{k-i} \left(Y_{t+i} - e_{t+i} - \sum_{j=1}^{i-1} (\alpha_{i-j} Y_{t+j} - \beta_{i-j} e_{t+j}) \right) \\ -(\alpha_{k-i} Y_{t+i} - \beta_{k-i} e_{t+i}) \end{array} \right) \\ &= e_{t+k} - \sum_{i=1}^{k-1} \left(\begin{array}{r} \left((\Lambda_{k-i} - \alpha_{k-i}) Y_{t+i} + (\beta_{k-i} - \Lambda_{k-i}) e_{t+i} \right) \\ -\Lambda_{k-i} \sum_{j=1}^{i-1} (\alpha_{i-j} Y_{t+j} - \beta_{i-j} e_{t+j}) \end{array} \right) . \end{aligned}$$

Gathering coefficients of Y_{t+i} gives, for $1 \leq i \leq k-1$,

$$\begin{aligned} & \left(\Lambda_{k-i} - \alpha_{k-i} - \Lambda_{k-i-1}\alpha_1 - \Lambda_{k-i-2}\alpha_2 - \ldots - \Lambda_1 \alpha_{k-i-1} \right) Y_{t+i} \\ = & - \left(\Lambda_{k-i} - \alpha_{k-i} - \sum_{j=1}^{k-1-i} \Lambda_{k-i-j} \alpha_j \right) Y_{t+i} , \end{aligned}$$

and since $\Lambda_0 = 1$, this means that the coefficient of Y_{t+i} is $-\left(\Lambda_{k-i} - \sum_{j=1}^{k-i} \Lambda_{k-i-j} \alpha_j \right)$;

but our definition of Λ_{k-i}, from 4.2, is exactly $\sum_{j=1}^{k-i} \alpha_j \Lambda_{k-i-j}$. So the coefficient of Y_{t+i}, for $1 \leq i < k-1$, is 0. The coefficient of Y_{t+k-1} is given simply as $(\Lambda_1 - \alpha_1) = -(\alpha_1 - \alpha_1) = 0$ too, and hence all the coefficients of Y_{t+i}, $1 \leq i \leq k-1$, are 0. Thus

$$e_t(k|D_t, V_0, \mathbf{W}_0) = e_{t+k} - \sum_{i=1}^{k-1} \left((\beta_{k-i} - \Lambda_{k-i}) e_{t+i} + \Lambda_{k-i} \sum_{j=1}^{i-1} \beta_{i-j} e_{t+j} \right) .$$

Then the coefficient of e_{t+i}, $1 \leq i \leq k-1$, is given by

$$\begin{aligned} & -(\beta_{k-i} - \Lambda_{k-i} + \Lambda_{k-i-1}\beta_1 + \Lambda_{k-i-2}\beta_2 + \ldots + \Lambda_1 \beta_{k-i-1}) \\ = \; & \Lambda_{k-i} - \beta_{k-i} - \sum_{j=1}^{k-i-1} \Lambda_{k-i-j}\beta_j = -\sum_{j=0}^{k-i} \Lambda_{k-i-j}\beta_j \; , \end{aligned}$$

since $\Lambda_0 = 1$ and we define $\beta_0 = -1$. Finally, the coefficient of e_{t+k-1} is $-(\beta_1 - \Lambda_1)$, with the e_{t+k} coefficient equal to 1. Thus we can write $e_t(k|D_t, V_0, \mathbf{W}_0)$ as given in 4.5, since this equation gives the required coefficients of e_{t+i}, $1 \leq i \leq k$.•

Example. Equation 4.5 in Lemma 4.2 enables us to calculate the theoretical step-ahead forecast error under any model. For instance,

$$e_t(1|D_t, V_0, \mathbf{W}_0) = \Lambda_0 \beta_0 e_{t+1} = e_{t+1} \tag{4.6}$$

(as by definition), and

$$\begin{aligned} e_t(2|D_t, V_0, \mathbf{W}_0) &= \Lambda_0\beta_1 e_{t+1} - \Lambda_1 \beta_0 e_{t+1} - \Lambda_0 \beta_0 e_{t+2} \\ &= (\alpha_1 - \beta_1) e_{t+1} + e_{t+2} \; . \end{aligned} \tag{4.7}$$

(Note the distinction, therefore, between $e_t(k|D_t, V_0, \mathbf{W}_0)$ and e_{t+k} in general.)

The straightforward addition of the two lemmas above into 4.4 now prove our final result of

Theorem 4.3: We can write the k^{th} step-ahead forecast error as

$$\hat{e}_t(k|D_t, V, \mathbf{W}) = e_t(k|D_t, V_0, \mathbf{W}_0) + \mathrm{E}[Y_{t+k}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+k}|D_t, V, \mathbf{W}]$$

$$= -\sum_{i=1}^{k}\left(\left(\sum_{j=0}^{k-i}\Lambda_j\beta_{k-i-j}\right)e_{t+i}\right)+\sum_{i=1}^{k}\Lambda_{k-i}\left(\sum_{j=i}^{p}\left(\beta_j-\hat{\beta}_j\left(\frac{1+\sum_{l=1}^{p}\beta_l\mathrm{B}^l}{1+\sum_{l=1}^{p}\hat{\beta}_l\mathrm{B}^l}\right)\right)e_{t+i-j}\right) . \tag{4.8}$$

Equation 4.8 has allowed us to express the k^{th} step-ahead forecast error under the specified model M in terms of the true k^{th} step-ahead forecast error, $e_t(k|D_t, V_0, \mathbf{W}_0)$ (under M_0), plus an extra term. This additional term of 4.3 implies that when we specify the variances within a constant TSDLM, we are introducing a *bias* into our forecast distribution, but one for which we can also calculate a precise algebraic form.

Before we do so, it is worth noting two points in relation to equation 4.8:

(i) as $V \to V_0$ and $\mathbf{W} \to \mathbf{W}_0$, the convergent form of $\mathbf{A}$ under M will tend to the true value $\mathbf{A}_0$, Hence $\mathbf{H} \to \mathbf{H}_0$ and its eigenvalues $\hat{\rho}_i \to \rho_i$; equivalently $\hat{\beta}_i \to \beta_i$. Thus the extra term in 4.8 becomes zero as we specify the true variance value, or in other words,

$$\hat{e}_t(k|D_t, V, \mathbf{W}) \underset{M_0}{\overset{M}{\to}} e_t(k|D_t, V_0, \mathbf{W}_0) ,$$

which is as we would expect.

(ii) In this second main term in 4.8 (arising from 4.3), the expression of the polynomial ratio in B, $\left(\frac{1+\sum_{l=1}^{p}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{p}\hat{\beta}_l \mathrm{B}^l}\right)$, will be (evidently) an expression of the form $1 + b_1\mathrm{B} + b_2\mathrm{B}^2 + \ldots$. Further, in this term as a whole, the summation over j is for $j = i, \ldots, p$ and so $i - j \leq 0$; hence we will always have

$$\left(\frac{1+\sum_{l=1}^{p}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{p}\hat{\beta}_l \mathrm{B}^l}\right) e_{t+i-j} = b_0 e_t + b_1 e_{t-1} + b_2 e_{t-2} + \ldots ,$$

a linear combination (with known constants b_i) in the true one step-ahead forecast errors at time $t-j$, for $j = 0, 1, \ldots$. The first term in 4.8, however, (arising from 4.5) will be a linear combination of e_{t+i} for $i = 1, \ldots, k$. Thus the two main terms in theorem 4.3 are indeed independent (once we have reached the limiting form for the constant TSDLM in question), a fact which we make use of in calculating the *exact* forecast distribution for two example models.

4.3 Example 1: the first-order polynomial model

4.3.1 Model definition

This is the simplest (and still non-trivial) DLM form, that we met briefly in section 3.4. It is a most widely used model that allows a practitioner to express the data evolution as being a locally constant underlying mean level, with a stochastic drift added in to allow longer-term changes in level.

The DLM is defined by the quadruple $\{1, 1, V, W\}$, giving the observation

equation

$$Y_t = \mu_t + v_t \text{ for } v_t \sim \mathcal{N}(0, V_t) ,$$

and the state equation

$$\mu_t = \mu_{t-1} + w_t \text{ where } w_t \sim \mathcal{N}(0, W_t) .$$

The term 'first-order polynomial' comes from seeing this latter equation for the underlying level of the series, μ_t, as $\mu(t + \delta t) = \mu(t)$ + higher order terms - it is the locally linear (or steady) model. The standard updating equations are obtained via the Kalman filter; for the posterior $(\mu_{t-1}|D_{t-1}) \sim \mathcal{N}(m_{t-1}, C_{t-1})$ we have prior

$$(\mu_t|D_{t-1}) \sim \mathcal{N}(m_{t-1}, R_t) , \quad R_t = C_{t-1} + W_t ,$$

with one step-ahead forecast

$$(Y_t|D_{t-1}) \sim \mathcal{N}(f_t, Q_t) , \quad f_t = m_{t-1} \text{ and } Q_t = R_t + V_t ,$$

(indeed, note that the k^{th} step-ahead forecast is $f_{t+k-1,k} = m_{t-1}$ for all $k \geq 1$), and posterior

$$\begin{aligned} (Y_t|D_t) &\sim \mathcal{N}(m_t, C_t) , \text{ where} \\ m_t &= m_{t-1} + A_t \hat{e}_t , \\ C_t &= A_t V_t , \end{aligned}$$

$$\hat{e}_t = Y_t - f_t \,,$$

$$\text{and } A_t = R_t/Q_t \,.$$

If, in addition, the observation and system variances are constant ($V_t = V$ and $W_t = W$), the model is known as the constant, closed (as the time series receives no external information) model. It is broadly used in sales forecasting and stock control, and is the model we study first since its relative simplicity lends great insight into the general use of equation 4.8 (and effectiveness of that utility) in calculating exact feedback estimates of V and W.

4.3.2 Limiting representation as ARIMA(0,1,1) process

Figure 4.1: 1st-order model, data (-), underlying level (..)

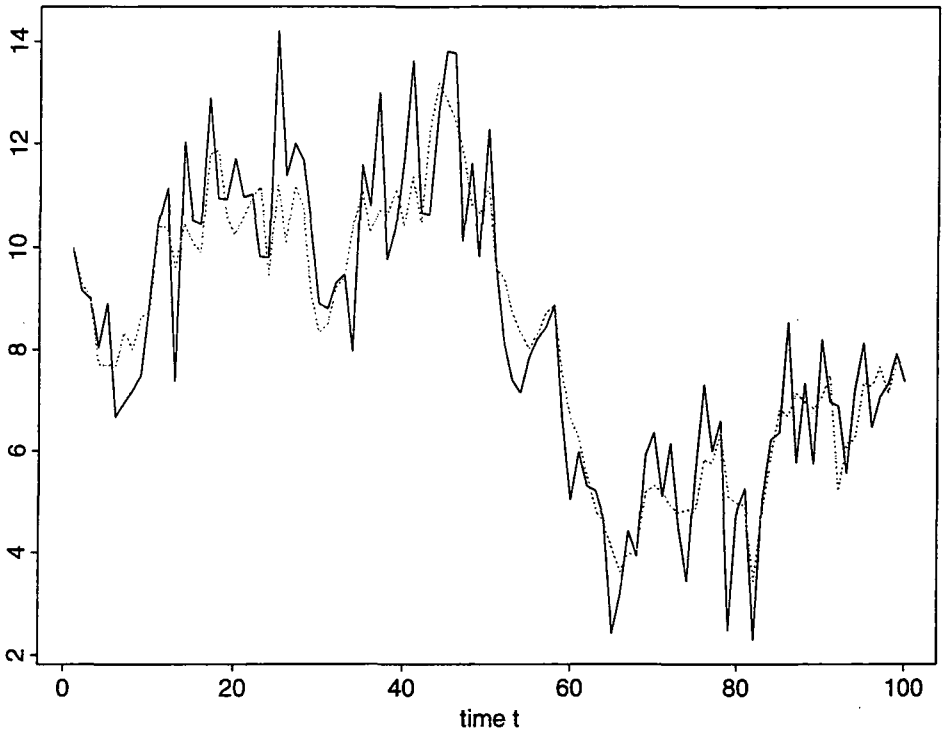


Figure 4.1 illustrates the first 100 points of a generated constant model of length 1000, with $\mu_0 = 10$, $V \equiv V_0 = 1$, and $W \equiv W_0 = 0.5$, together with the evolution of the underlying level μ_t of the series over this time (the first 200 points of this data set appears as Series 3 in the Appendix). The behaviour of the series, with its 'dependence on level', is rather akin to that of the Box-Jenkins [3] ARIMA(0,1,1) process. There is good cause for this - the limiting form for the constant model is indeed exactly such a non-stationary Box-Jenkins process; for it easily shown (see, for instance, West and Harrison [36]) that as the closed model updates upon receiving new data, the convergent form for the adaptive coefficient $A_t = R_t/Q_t$ is reached monotonically, and often rapidly, and is given as

$$A = r(\sqrt{1 + 4/r} - 1)/2 \ ,$$

where $r = W/V$ is known as the signal-to-noise ratio. Then, from the updating equations above, remembering that we are taking constant $V_t = V$ and $W_t = W$, we also have the following convergences:

$$\begin{aligned} C_t \to C &= AV \\ R_t \to R &= AV/(1-A) \\ \text{and } Q_t \to Q &= R + V = V/(1-A) \ . \end{aligned}$$

So with $\hat{e}_t = Y_t - m_{t-1}$ and $m_t = m_{t-1} + A_t\hat{e}_t$,

$$Y_t - Y_{t-1} = \hat{e}_t + m_{t-1} - \hat{e}_{t-1} - m_{t-2}$$

$$
\begin{aligned}
&= \hat{e}_t - (1 - A_{t-1})\hat{e}_{t-1} \\
\Rightarrow \lim_{t\to\infty}(Y_t - Y_{t-1}) &= \hat{e}_t - (1 - A)\hat{e}_{t-1} \\
&= \hat{e}_t - \delta\hat{e}_{t-1} \quad \text{where} \quad 0 < \delta = 1 - A < 1 \ ,
\end{aligned}
$$

and with $\mathrm{Var}(\hat{e}_t|D_{t-1}) = \mathrm{Var}(Y_t|D_{t-1}) = Q_t$ converging as above to Q, the limiting form is

$$
Y_t = Y_{t-1} + \hat{e}_t - \delta\hat{e}_{t-1} \ , \quad \hat{e}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, Q) \ .
$$

This is equivalent to the Box-Jenkins ARIMA(0,1,1) model of

$$
(1 - \mathrm{B})Y_t = (1 + \theta\mathrm{B})a_t \ , \quad |\theta| < 1, \ a_t \sim \mathcal{N}(0, \sigma_a^2) \ ,
$$

where, again, B is the backward shift operator such that $\mathrm{B}a_t = a_{t-1}$.

4.3.3 Effects of model misspecification

Before applying theorem 4.3 to the first-order polynomial model, it is worth noting that the aim of the following sections (namely estimation of V and W, and, later in the chapter with respect to the second-order model, estimation of V and $\mathbf{W}$) can also be reached through a maximum likelihood estimation procedure, although this approach has not been pursued here. Through consideration of the first difference of the data series $\{Y_t\}$ in the first-order model,

and considering instead the series $\{z_t\}$ (for $z_t = Y_t - Y_{t-1}$), we notice that

$$z_t = \mu_t + v_t - \mu_{t-1} - v_{t-1} = w_t + v_t - v_{t-1} \ ,$$

giving a relation solely in terms of the error sequences $\{v_t\}$ and $\{w_t\}$. Hence by looking at the covariance structure of the zero-mean $\{z_t\}$ series (which will be expressible purely in terms of the variances V and W), we can use a maximum likelihood estimation procedure to solve for both V and W.

To return to the above limiting form for the model, however, we can also apply the results of theorem 4.3. We have that $\mathbf{H_0} = (\mathbf{I} - \mathbf{A_0}\mathbf{F}^T)\mathbf{G} \Rightarrow H_0 = (1 - A_0)$, and so the eigenvalue of H_0 is simply $\rho = \delta_0 = 1 - A_0$, with $\lambda = 1$. Thus $\alpha_1 = 1$ and $\beta_1 = A_0 - 1$ (giving, incidentally, from equation 2.11 that $Y_t = \alpha_1 Y_{t-1} + e_t + \beta_1 e_{t-1}$ is indeed the same Box-Jenkins ARIMA(0,1,1) representation). Then equations 4.6 and 4.8 give us

$$\begin{aligned}
\hat{e}_t(1|D_t, V, W) &= e_{t+1} + \left(\beta_1 - \hat{\beta}_1\left(\frac{1 + \sum_{l=1}^{p} \beta_l \mathrm{B}^l}{1 + \sum_{l=1}^{p} \hat{\beta}_l \mathrm{B}^l}\right)\right) e_t \\
&= e_{t+1} + \left(\beta_1 - \hat{\beta}_1(1+\beta_1\mathrm{B})(1 - \hat{\beta}_1\mathrm{B} + \hat{\beta}_1^2\mathrm{B}^2 - \hat{\beta}_1^3\mathrm{B}^3 + \ldots)\right) e_t \\
&= e_{t+1} + \left(\beta_1 - \hat{\beta}_1(1 + (\beta_1 - \hat{\beta}_1)\mathrm{B} + (\hat{\beta}_1^2 - \beta_1\hat{\beta}_1)\mathrm{B}^2 + (\beta_1\hat{\beta}_1^2 - \hat{\beta}_1^3)\mathrm{B}^3 + \ldots)\right) e_t \ .
\end{aligned}$$

Thus - noting that the $e_t(1|D_t, V_0, W_0) = e_{t+1}$ term is indeed independent of the rest of this equation - we have that

$$\mathrm{Var}(\hat{e}_t(1|D_t, V, W)) \ = \ Q_0 + \mathrm{Var}\left((\beta_1 - \hat{\beta}_1)(1 - \hat{\beta}_1\mathrm{B} + \hat{\beta}_1^2\mathrm{B}^2 - \hat{\beta}_1^3\mathrm{B}^3 + \ldots)e_t\right)$$

$$= Q_0 + (\beta_1 - \hat{\beta}_1)^2 \frac{Q_0}{1 - \hat{\beta}_1^2} .$$

So with $\beta_1 = -\delta_0$ as above, this has produced

$$\mathrm{Var}(\hat{e}_t(1|D_t, V, W)) = Q_0 \left(1 + \frac{(\delta - \delta_0)^2}{1 - \delta^2}\right) . \tag{4.9}$$

We can extend this to the k^{th} step-ahead forecast error $\hat{e}_t(k|D_t, V, W)$. From equation 4.2, since $\hat{\alpha}_1 = \alpha_1 = 1$ and $\alpha_k = 0$ for $k > 1$, we have $\Lambda_k = 1$ for all $k = 0, 1, \ldots$ and so together with $\beta_0 = -1$, $\beta_1 = \delta_0$ and $\beta_k = 0$ for $k > 1$, this gives from equation 4.5 that

$$\begin{aligned} e_t(k|D_t, V_0, W_0) &= -(\delta_0 - 1)e_{t+1} - (\delta_0 - 1)e_{t+2} - \ldots - (\delta_0 - 1)e_{t+k-1} + e_{t+k} \\ &= e_{t+k} + (1 - \delta_0)(e_{t+k-1} + \ldots + e_{t+1}) . \end{aligned} \tag{4.10}$$

Then 4.3 yields

$$\begin{aligned} \mathrm{E}[Y_{t+k}|D_t, V_0, W_0] - \mathrm{E}[Y_{t+k}|D_t, V, W] &= \left(\beta_1 - \hat{\beta}_1 \left(\frac{1 + \beta_1 \mathrm{B}}{1 + \hat{\beta}_1 \mathrm{B}}\right)\right) e_t \\ &= \left(\delta_0 - \delta \left(\frac{1 + \delta_0 \mathrm{B}}{1 + \delta \mathrm{B}}\right)\right) e_t \\ &= \left(\frac{\delta_0 - \delta}{1 + \delta \mathrm{B}}\right) e_t , \end{aligned} \tag{4.11}$$

independent of k. (Both these results are cited in similar forms for the ARIMA(0,1,1) process in Box and Jenkins, pps. 267-268.)

So again noting that equations 4.10 and 4.11 are independent, we can find

from 4.8 that

$$\begin{aligned}\mathrm{Var}(\hat{e}_t(k|D_t,V,W)) &= \mathrm{Var}(e_{t+k}+(1-\delta_0)(e_{t+k-1}+\ldots+e_{t+1}))+\mathrm{Var}\left(\left(\frac{\delta_0-\delta}{1+\delta \mathrm{B}}\right)e_t\right)\\ &= Q_0(1+(1-\delta_0)^2(k-1))+(\delta-\delta_0)^2\mathrm{Var}(e_t-\delta e_{t-1}+\delta^2 e_{t-2}-\ldots)\\ &= Q_0(1+(1-\delta_0)^2(k-1))+(\delta-\delta_0)^2\frac{Q_0}{1-\delta^2}\,. \qquad (4.12)\end{aligned}$$

Equations 4.9 and 4.12 are valuable results. Not only do they confirm that as $M \to M_0$ (so that $V \to V_0$, $W \to W_0$ and $A \to A_0$, $\delta \to \delta_0$), we have the intuitively required result of $\mathrm{Var}(\hat{e}_t(1|D_t,V,W)) \underset{M_0}{\overset{M}{\to}} Q_0 = \mathrm{Var}(e_t(1|D_t,V_0,W_0))$, and, more generally,

$$\mathrm{Var}(\hat{e}_t(k|D_t,V,W)) \underset{M_0}{\overset{M}{\to}} \mathrm{Var}(e_t(k|D_t,V_0,W_0))\,,$$

but we can also see the effects of misspecifying the signal-to-noise ratio r as this misspecification becomes more and more severe:

(i) recalling that $A = r(\sqrt{1+4/r}-1)/2$, then

$$\begin{aligned} A &= (\sqrt{r^2+4r}-r)/2 \\ &\underset{0}{\overset{r}{\to}} 0\,, \end{aligned}$$

$$\text{so that } \delta = 1-A \to 1 \text{ as } r \to 0\,.$$

Thus, from equation 4.12

$$\mathrm{Var}(\hat{e}_t(k|D_t,V,W)) \underset{0}{\overset{r}{\to}} \infty\,,$$

and so drastically *underestimating* the ratio $r = W/V$ (in other words taking W too small in relation to V), leads to the data effectively being ignored and, due to the 'dependence on level' of the process, which results in large meanderings of the data away from μ_0, the variance of the k^{th} step-ahead forecast errors tends to infinity.

(ii) Now, noting that $\sqrt{1+4/r}$ can be written as $1 + \frac{2}{r} - \frac{2}{r^2} + \frac{4}{r^3} - \ldots$, we have

$$\begin{aligned} A &= (2 - \frac{2}{r} + \frac{4}{r^2} - \ldots)/2 \to 1 \text{ as } r \to \infty \\ \Rightarrow \delta = 1 - A &\underset{\infty}{\overset{r}{\to}} 0 . \end{aligned}$$

So, from equation 4.12,

$$\begin{aligned} \text{Var}(\hat{e}_t(k|D_t, V, W)) &\underset{\infty}{\overset{r}{\to}} Q_0(1 + (1-\delta_0)^2(k-1)) + \delta_0^2 Q_0 \\ &= \text{Var}(e_t(k|D_t, V_0, W_0)) + \delta_0 V_0 . \end{aligned}$$

This time, as we *overestimate* the signal-to-noise ratio by setting W too large with respect to V, the k^{th} step-ahead forecast error sequence variance is finite, and inflated by a (potentially small) value of $\delta_0 V_0$ - with $V_0 = 1$, $W_0 = 1/2$ this value is $1/2$. As West and Harrison state, this is a far more desirable situation to be in (even more so for reasons we shall see presently), and so in cases of uncertainty it is always 'best' (in terms of forecasting performance, at least) to overestimate W with respect to V.

The full effects of the misspecification of r on the inflation in the limiting

value of $\mathrm{Var}(\hat{e}_t(k|D_t, V, W))$, for various true values of $r_0 = W_0/V_0$, are shown in Table (4.1) below.

Table 4.1: effects of r misspecification on $\mathrm{Var}(\hat{e}_t(k|D_t, V, W))$.

V_0	r_0	δ_0	r	δ	A_0Q_0	Inflation in $\mathrm{Var}(\hat{e}_t(k\|D_t, V, W))$ of $(\delta-\delta_0)^2\frac{Q_0}{1-\delta^2}$	$\delta_0 V_0$
1	1/2	1/2	0.05	0.80	1	0.500	
			5×10^{-3}	0.93		2.827	
			5×10^{-4}	0.98		10.444	
			1	0.38		0.033	
			10	0.084		0.349	
			100	9.8×10^{-3}		0.481	1/2
2	1/4	0.610	0.05	0.80	1.281	0.330	
			5×10^{-3}	0.93		2.582	
			5×10^{-4}	0.98		10.174	
			1	0.38		0.199	
			10	0.084		0.913	
			100	9.8×10^{-3}		1.180	1.219

(The A_0Q_0 column has been inserted for reference in equation 4.13 below).

4.3.4 Estimation of V and W

Now we return to the task in hand of calculating V_0 and W_0. From the independence of equations 4.10 and 4.11, we have

$$\begin{aligned}\mathrm{Cov}(Y_{t+j}, Y_{t+k}|D_t, V, W) &= \mathrm{E}[\hat{e}_t(j|D_t, V, W)\hat{e}_t(k|D_t, V, W)] \\ &= \mathrm{E}[e_t(j|D_t, V_0, W_0)e_t(k|D_t, V_0, W_0)] + \mathrm{Var}\left(\left(\frac{\delta_0-\delta}{1+\delta\mathrm{B}}\right)e_t\right).\end{aligned}$$

Specifically, from equations 4.6 and 4.7,

$$\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, W) = \mathrm{E}[e_{t+1}(e_{t+2}+(1-\delta_0)e_{t+1})|D_t, V_0, W_0] + \mathrm{Var}\left(\left(\frac{\delta_0-\delta}{1+\delta\mathrm{B}}\right)e_t\right)$$

$$= (1-\delta_0)Q_0 + (\delta-\delta_0)^2\frac{Q_0}{1-\delta^2}$$

$$= A_0Q_0 + (\delta-\delta_0)^2\frac{Q_0}{1-\delta^2}\ . \tag{4.13}$$

Again, this is a valuable result. It confirms the intuitive limiting result of $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, W) \underset{M_0}{\overset{M}{\rightarrow}} A_0Q_0 = \mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V_0, W_0)$ (from equation 3.5, $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V_0, W_0) = C_0 + W_0 = R_0 = A_0Q_0$), as well as - far more importantly - enabling us to calculate V_0 from knowledge of $\mathrm{Var}(\hat{e}_t(1|D_t, V, W))$ and $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, W)$. For, by noticing that the inflations in equations 4.13 and 4.9 are identical, subtracting these two produces

$$\mathrm{Var}(\hat{e}_t(1|D_t, V, W)) - \mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, W) = Q_0 - A_0Q_0 = V_0\ . \tag{4.14}$$

So we can estimate V_0 by simply calculating the sample one step-ahead forecast error variance $S_1^2 = S^2(\hat{e}_t(1|D_t, V, W))$, and the sample lag-one covariance $\widehat{C}_{12}$, and subtracting them. This estimate is not only stable, since now we are accounting for *any* variance misspecification in the model M, but is also highly accurate; its precision depends only upon the accuracy of the two sample estimates S_1^2 and $\widehat{C}_{12}$, whose convergence properties we will discuss presently.

Extending our calculations to estimation of W, substituting $\widehat{C}_{12}$ in 4.13 yields

$$\widehat{C}_{12} = A_0Q_0 + (\delta-\delta_0)^2\frac{Q_0}{1-\delta^2}\ ,$$

$$\Rightarrow (1-\delta_0)\frac{V_0}{\delta_0} + (\delta-\delta_0)^2\frac{V_0}{\delta_0(1-\delta^2)} - \widehat{C}_{12} = 0$$

$$\Rightarrow V_0\delta_0^2 + (V_0(\delta^2-2\delta-1) - (1-\delta^2)\widehat{C}_{12})\delta_0 + V_0 = 0\ ,$$

and $A_0 = r_0\left(\sqrt{1+4/r_0}-1\right)/2$ gives $W_0 = V_0(1-\delta_0)^2/\delta_0$, so that

$$\begin{aligned}
\delta_0 W_0 &= V_0 - 2V_0\delta_0 + V_0\delta_0^2 \\
&= V_0 - 2V_0\delta_0 + \delta_0((1-\delta^2)\widehat{C}_{12} - V_0(\delta^2-2\delta-1)) - V_0 \\
\Rightarrow W_0 &= (1-\delta^2)\widehat{C}_{12} - V_0(\delta-1)^2\ .
\end{aligned}$$

This result, from equation 4.14 and by substituting the sample values S_1^2 and $\widehat{C}_{12}$ for $\mathrm{Var}(\hat{e}_t(1|D_t,V,W))$ and $\mathrm{Cov}(Y_{t+1},Y_{t+2}|D_t,V,W)$ respectively, then gives

$$W_0 = 2\widehat{C}_{12}(1-\delta) - S_1^2(\delta-1)^2\ . \tag{4.15}$$

So Theorem 4.3 has indeed been sufficient to produce *exact* solutions for both V_0 and W_0, assuming that both the sample estimates $\widehat{C}_{12}$ and S_1^2 are accurate. This is the 'big if', of course - how rapidly the sample covariance of $\frac{1}{n-1}\sum_{i=2}^{n}(Y_{i-1} - f_{i-1})(Y_i - f_{i,2})$ converges to its true value of $\widehat{C}_{12}$, and how the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}((Y_i - f_i) - \bar{\hat{e}}_1)^2$, for $\bar{\hat{e}}_1 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_i)$, converges to its true value of S_1^2.

4.3.5 Convergence properties of $\widehat{V}$ and $\widehat{W}$

We tackle this issue by further calculation of the variances of both $\widehat{C}_{12}$ and S_1^2. The estimate $\widehat{C}_{12}$ is calculated by taking samples $(Y_{t+1}-m_t)(Y_{t+2}-m_t)$, of a ran-

dom variable C_{12}, say, which has expected value $\mathrm{E}[C_{12}] = \mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, W)$. Then, by noting that

$$\begin{aligned} C_{12} &= \hat{e}_t(2|D_t, V, W)\hat{e}_t(1|D_t, V, W) \\ &= \left((e_{t+2}+(1-\delta_0)e_{t+1})+\left(\frac{\delta-\delta_0}{1-\delta B}\right)e_t\right)\left(e_{t+1}+\left(\frac{\delta-\delta_0}{1-\delta B}\right)e_t\right) , \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(C_{12}) = \mathrm{Var}\Bigg(& e_{t+2}e_{t+1} + (1-\delta_0)e_{t+1}^2 + (2-\delta_0)\left(\frac{\delta-\delta_0}{1-\delta B}\right)e_te_{t+1} \\ &+ \left(\frac{\delta-\delta_0}{1-\delta B}\right)e_te_{t+2} + \left(\frac{\delta-\delta_0}{1-\delta B}\right)^2 e_t^2\Bigg) . \end{aligned} \tag{4.16}$$

Now, $(e_t|D_t)$, $(e_{t+1}|D_t)$, $(e_{t+2}|D_t)$ are all $\overset{iid}{\sim}\mathcal{N}(0, Q_0)$. Thus

$$\mathrm{Var}(e_{t+2}e_{t+1}) = \mathrm{E}[e_{t+2}^2]\mathrm{E}[e_{t+1}^2] = Q_0^2 , \tag{4.17}$$

and

$$\mathrm{Var}((1-\delta_0)e_{t+1}^2) = (1-\delta_0)^2(\mathrm{E}[e_{t+1}^4] - Q_0^2)$$

in which $\mathrm{E}[e_{t+1}^4]$ is found from the fourth derivative of the moment generating function of e_{t+1}, $m(u) = e^{Q_0\frac{u^2}{2}}$, evaluated at 0 - this gives $\mathrm{E}[e_{t+1}^4] = 3Q_0^2$, whereupon

$$\mathrm{Var}((1-\delta_0)e_{t+1}^2) = 2(1-\delta_0)^2Q_0^2 . \tag{4.18}$$

Further,

$$\mathrm{Var}((2-\delta_0)\left(\frac{\delta-\delta_0}{1-\delta B}\right)e_te_{t+1}) = (2-\delta_0)^2(\delta-\delta_0)^2\mathrm{Var}((e_t+\delta e_{t-1}+\delta^2e_{t-2}+\ldots)e_{t+1})$$

whereupon we must deal with the all covariance terms that arise... However, all of the $\mathrm{E}[e_{t-i}e_{t-j}e_{t+1}^2]$ and $\mathrm{E}[e_{t+1}e_{t-k}]$ are zero due to the independence of all the functions of the e's, and so these covariances are all zero too, yielding

$$\begin{aligned}\mathrm{Var}\left((2-\delta_0)\left(\frac{\delta-\delta_0}{1-\delta B}\right)e_te_{t+1}\right) &= (2-\delta_0)^2(\delta-\delta_0)^2(\mathrm{Var}(e_te_{t+1})+\delta^2\mathrm{Var}(e_{t-1}e_{t+1})+\ldots) \\ &= (2-\delta_0)^2(\delta-\delta_0)^2\frac{Q_0^2}{1-\delta^2}\,. \end{aligned} \tag{4.19}$$

Similarly,

$$\mathrm{Var}\left(\left(\frac{\delta-\delta_0}{1-\delta B}\right)e_te_{t+2}\right) = (\delta-\delta_0)^2\frac{Q_0^2}{1-\delta^2}\,, \tag{4.20}$$

but the final variance term of $\mathrm{Var}(\left(\frac{\delta-\delta_0}{1-\delta B}\right)^2 e_t^2)$ is rather more complicated; we have

$$\begin{aligned}\mathrm{Var}\left((\delta-\delta_0)^2\left(\frac{e_t}{1-\delta B}\right)^2\right) &= (\delta-\delta_0)^4\mathrm{Var}((e_t+\delta e_{t-1}+\delta^2e_{t-2}+\ldots)^2) \\ &= (\delta-\delta_0)^4\mathrm{Var}(e_t(e_t+\delta e_{t-1}+\ldots)+\delta e_{t-1}(e_t+\delta e_{t-1}+\ldots) \\ &\quad +\delta^2e_{t-2}(e_t+\delta e_{t-1}+\ldots)+\ldots)\,.\end{aligned}$$

But again, all the covariance terms are zero, since none of the e_{t-i}^2 terms are repeated - this gives

$$\begin{aligned}\mathrm{Var}\left((\delta-\delta_0)^2\left(\frac{e_t}{1-\delta B}\right)^2\right) &= (\delta-\delta_0)^4(2Q_0^2+\delta^2Q_0^2+\delta^4Q_0^2+\ldots+\delta^2Q_0^2+2\delta^4Q_0^2+\delta^6Q_0^2+\ldots \\ &\quad +\delta^4Q_0^2+\delta^6Q_0^2+2\delta^8Q_0^2+\delta^{10}Q_0^2+\ldots \\ &= (\delta-\delta_0)^4Q_0^2\left(\left(1+\frac{1}{1-\delta^2}\right)+\delta^2\left(\delta^2+\frac{1}{1-\delta^2}\right)+\delta^4\left(\delta^4+\frac{1}{1-\delta^2}\right)+\ldots\right)\end{aligned}$$

$$= (\delta - \delta_0)^4 Q_0^2 \left(\frac{1}{1-\delta^4} + \frac{1}{(1-\delta^2)^2} \right) . \tag{4.21}$$

So finally, referring back to equation 4.16 above and again remarking that all the covariance terms within it are zero, we reach (through summation of equations 4.17, 4.18, 4.19, 4.20 and 4.21)

$$\begin{aligned} \mathrm{Var}(C_{12}) &= Q_0^2 + 2(1-\delta_0)^2 Q_0^2 + (2-\delta_0)^2 (\delta - \delta_0)^2 \frac{Q_0^2}{1-\delta^2} \\ &\quad + (\delta - \delta_0)^2 \frac{Q_0^2}{1-\delta^2} + (\delta - \delta_0)^4 Q_0^2 \left(\frac{1}{1-\delta^4} + \frac{1}{(1-\delta^2)^2} \right) \\ \Rightarrow \mathrm{Var}(C_{12}) &= Q_0^2 \left(1 + 2(1-\delta_0)^2 + (\delta - \delta_0)^2 \left(\frac{(2-\delta_0)^2 + 1}{1-\delta^2} + \frac{(\delta-\delta_0)^2}{1-\delta^4} + \frac{(\delta-\delta_0)^2}{(1-\delta^2)^2} \right) \right) . \end{aligned}$$

(This is the theoretical variance of $\mathrm{Cov}(Y_{t+1}, Y_{t+2} | D_t, V, W)$, of course, whereas we are interested in the theoretical variance of the sample, $\mathrm{Var}(\widehat{C}_{12})$ - this comes from the Central Limit Theorem as

$$\mathrm{Var}(\widehat{C}_{12}) = \frac{1}{n-1} \mathrm{Var}(C_{12}) \; .)$$

Under the correct model specification of $M = M_0$, therefore, the convergence of $\widehat{C}_{12}$ to $\mathrm{E}[C_{12}] = \mathrm{Cov}(Y_{t+1}, Y_{t+2} | D_t, V, W)$ is indeed relatively rapid, with $\mathrm{Var}(\widehat{C}_{12}) = \frac{Q_0^2}{n-1}(1+2(1-\delta_0)^2)$; taking $V_0 = 1$, $W_0 = 1/2$ gives $\delta_0 = 1/2$, $Q_0 = 2$, and so $\mathrm{Var}(\widehat{C}_{12}) = 6/(n-1)$. Additionally, we can again see the effects of misspecification of r:

(i) as $\delta \to 0$ $(r \to \infty)$, $\mathrm{Var}(C_{12}) \;\to\; Q_0^2(1 + 2(1-\delta_0)^2 + \delta_0^2((2-\delta_0)^2 + 1 + 2\delta_0^2)$

$$= \quad Q_0^2(3 - 4\delta_0 + 7\delta_0^2 - 4\delta_0^3 + 3\delta_0^4) \ ,$$

which with $V_0 = 1,\ W_0 = 1/2$ is 9.75. Hence as $r \to \infty$, $\mathrm{Var}(\widehat{C}_{12}) \to 9.75/(n-1)$ with this model specification; in general, since $0 < \delta_0 < 1$, $\mathrm{Var}(\widehat{C}_{12})$ will have a finite and relatively small upper bound of $5Q_0^2/(n-1)$, so convergence of $\widehat{C}_{12}$ to $\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, W)$ will always be rapid through overestimating $r = W/V$. However:

(ii) as $\delta \to 1$ $(r \to 0)$, $\mathrm{Var}(C_{12}) \to \infty$ and so underestimation of r not only causes the estimate $\widehat{W}$ to become more and more biassed, but convergence of $\widehat{C}_{12}$ also slows considerably, again showing the value of *overestimation* of the signal-to-noise ratio.

Now consideration turns to the sample variance of the one step-ahead forecast errors, $S_1^2 = \mathrm{S}^2(\hat{e}_t(1|D_t, V, W)) \to Q_0 + (\delta - \delta_0)^2 \frac{Q_0}{1-\delta^2}$. The rate of convergence of this sample variance will depend upon its own variance. If we take

$$\mathrm{X} = \hat{e}_t(1|D_t, V, W) \quad = \quad e_{t+1} + \left(\frac{\delta - \delta_0}{1 - \delta B}\right) e_t \sim \mathcal{N}(0, \sigma_X^2) \ ,$$

$$\text{where} \quad \sigma_X^2 \quad = \quad Q_0 + (\delta - \delta_0)^2 \frac{Q_0}{1 - \delta^2} \ ,$$

$$\text{then} \quad S_1^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i^2 \quad \Rightarrow \quad n\frac{S_1^2}{\sigma_X^2} \sim \chi_{n-1}^2$$

$$\text{whence} \quad \mathrm{Var}(S_1^2) \quad = \quad \frac{2(n-1)}{n^2}(\sigma_X^2)^2$$

$$= \quad \frac{2(n-1)}{n^2}\left(Q_0 + (\delta - \delta_0)^2 \frac{Q_0}{1 - \delta^2}\right)^2 \ .$$

With $\delta = \delta_0$, $\mathrm{Var}(S_1^2) = \frac{2(n-1)}{n^2} Q_0^2$, and so convergence of S_1^2 is fast, as is the case

when we overestimate r ($\delta \to 0$), when $\text{Var}(S_1^2) \to \frac{2(n-1)}{n^2} Q_0^2 (1+\delta_0^2)^2$. Again, however, by underestimating r ($\delta \to 1$) this variance ultimately tends to infinity and convergence of S_1^2 is slowed.

4.3.6 Simulation results

We have seen many theoretical results thus far, and these are best understood through practical illustration; we firstly take a simulated constant model series of length 1000 with $V_0 = 1$, $W_0 = 1/2$, so that $A_0 = \delta_0 = 1/2$, $Q_0 = 2$, and show the results of fitting a constant model with correct and various incorrect specifications of $r = W/V$, for both $n = 100$ and then the full series ($n = 1000$), in Table 4.2:

Table 4.2: results for various fitted constant models on simulated series (Series 3 in the Appendix), up to lengths $n = 100$ and $n = 1000$, with $V_0 = 1,\ W_0 = 1/2$

			C_{12}		$\mathrm{Var}(\hat{e}_t(1\|D_t,\delta))$		Feedback	Feedback	$\mathrm{Var}(C_{12})$		Theor-
			Sample value	Theor- etical	Sample value	Theor- etical	estimate of V_0,	estimate of W_0,	Sample value	Theor- etical	etical $\mathrm{Var}(S_1^2)$
V	W	δ	$\hat{C}_{12}$	value	S_1^2	value	$S_1^2 - \hat{C}_{12}$	eq. 4.15		value	$(\times n)$
Up to length $n = 100$											
1.0	0.5	0.5	1.163	1	2.374	2	1.211	0.569	6.291	6	7.92
1.0	0.05	0.8	1.604	1.5	2.826	2.5	1.222	0.528	9.352	9.555	12.38
0.1	0.005	0.8	1.595	1.5	2.820	2.5	1.225	0.525	9.376	9.555	12.38
1.0	0.005	0.932	3.268	3.827	4.402	4.827	1.134	0.426	32.21	32.94	46.13
1.0	0.0005	0.978	5.484	11.44	6.077	12.44	0.593	0.240	72.63	185.4	306.6
1.0	1	0.382	1.210	1.033	2.420	2.033	1.210	0.572	6.549	6.214	8.184
1.0	10	0.084	1.581	1.349	2.800	2.349	1.219	0.546	9.558	8.508	10.93
10	100	0.084	1.581	1.349	2.800	2.349	1.219	0.546	9.558	8.508	10.93
1.0	100	0.010	1.733	1.481	2.958	2.481	1.225	0.532	11.08	9.586	12.19
Up to length $n = 1000$											
1.0	0.5	0.5	0.921	1	1.954	2	1.033	0.432	5.252	6	7.992
1.0	0.05	0.8	1.281	1.5	2.320	2.5	1.039	0.420	7.604	9.555	12.49
0.1	0.005	0.8	1.280	1.5	2.321	2.5	1.041	0.419	7.606	9.555	12.49
1.0	0.005	0.932	3.155	3.827	4.196	4.827	1.041	0.411	25.83	32.94	46.55
1.0	0.0005	0.978	7.262	11.44	8.331	12.44	1.070	0.317	119.7	185.4	309.2
1.0	1	0.382	0.972	1.033	2.003	2.033	1.031	0.437	5.613	6.214	8.258
1.0	10	0.084	1.309	1.349	2.341	2.349	1.032	0.434	8.226	8.508	11.02
10	100	0.084	1.309	1.349	2.343	2.349	1.034	0.433	8.226	8.508	11.02
1.0	100	0.010	1.443	1.481	2.478	2.481	1.034	0.429	9.399	9.586	12.30

The third and eighth lines in both halves of the table have been put in to illustrate how it is only the signal-to-noise *ratio* $r = W/V$ which is of importance in the calculations, and so for the remainder of the fitted models V has been left at a (computationally) convenient value of 1. There are also several other features worthy of special note:

(i) the convergences of $\hat{C}_{12}$ and S_1^2 behave almost exactly as predicted - they are indeed fairly slow when r is underspecified (more so for $n = 100$), but for r overestimated both of these estimates are remarkably accurate, as is the behaviour of the sample variance of $\hat{C}_{12}$.

(ii) We can also see how both $\widehat{C}_{12}$ and S_1^2 are indeed affected in a very similar manner (remember that the bias in both of these estimates is theoretically identical), and so the overall effect on the calculated estimate of V_0 - just the difference in the two estimates - is small, (negligible for $n = 1000$), as is the influence on the feedback estimate of W_0. In fact, even for gross misspecification of r (by factors of 200 and 1000 respectively), the resulting calculated estimates for V_0 when $n = 1000$ are out by only 3.4% and 7%, and those for W_0, although in error by slightly more, are so far removed from the initial model specifications of W that the forecaster would soon be moved into rethinking them! (Leading, of course, into a significantly more accurate specification area of r, where convergence would then be rapidly achieved).

(iii) There appears to be some evidence of bias in the variance estimates, nonetheless; it is more apparent in the estimates of V_0 at $n = 100$, and in those for W_0 at n=1000. This bias is most probably due to the non-linearity of equations 4.9, 4.13 and 4.15 with respect to the these estimates in terms of the relevant sample variance and covariance estimates, but is mostly negligible, especially in the most severe model misspecifications when compared to the scale of error in the original variance specifications.

4.3.7 Comparison with sub-optimal filtering

The closest that previous work has come to combatting the problem of divergence of estimates, due to sub-optimal filtering in the steady model, has been by Cantarelis and Johnston [7] in 1983. They tackle on-line variance estimation in this model via a maximum likelihood approach, from looking at the likelihood

$L_t \propto p(Y_t|D_{t-1}, V, W) = \frac{1}{\sqrt{Q_t}} e^{\{\frac{-1}{2Q_t}(\hat{e}_t^2)\}}$, where $\hat{e}_t = Y_t - f_t$, the one step-ahead forecast error in the usual notation. Taking the limiting form of $Q_t \rightarrow Q$, and maximising $\mathrm{E}^* = \mathrm{E}[\ln(L_t)]$ with respect to V and A (since we know the variance of $\hat{e}_t$) then yields the MLE's of

$$\begin{aligned} V^* &= (1-A)\mathrm{E}[\hat{e}_t^2] \\ \text{and } A^* &= \frac{r}{2}\left(\sqrt{1+\frac{4}{r}} - 1\right) . \end{aligned}$$

As it stands, this is nothing more, of course, than a sub-optimal filtering technique again, since we are merely inverting existing limiting equations from the Kalman Filter. Cantarelis and Johnston try and minimise the risk of the potential resulting divergence by adopting a multi-process class I approach, where it is assumed that j alternative models, $M^{(k)}$ for $k = 1, \ldots, j$, will, between them, adequately describe the correct model. Each model has an uncertainty $p_t^{(k)}$ associated with it at time t, updated via Bayes' theorem, as well as a particular variance ratio of $r^{(k)}$. Practically, to provide a black-box method for setting up the approach, a total of $j = 8$ models are generally used, with $r^{(1)} = 1$, $r^{(k)} = r^{(k-1)}/2$ for $k = 2, \ldots 8$ (so that $r^{(8)} = 1/128$, a reasonably wide $r^{(k)}$ range), and uninformative priors $p_0^{(k)} = 1/8$, $C_0^{(k)} = 100V_0^{(k)}$ (for initial V estimate of $V_0^{(k)}$ in the k^{th} model). At each variance updating stage (taken to be every ten time points), the overall posteriors for $\widehat{V}_t$, $\widehat{W}_t$, m_t and C_t are calculated using a probability weighting combination of the relevant parameters from each model $M^{(k)}$.

This multi-process method has two obvious drawbacks. The first is that we

may still face divergence of our solutions, due to the chosen set of $r^{(k)}$'s not containing the true value r_0; the second that we evidently face a huge increase in computational complexity and time by having to carry so many alternative models, updating each one and re-estimating our posteriors at each stage via probability weighting. We have no prior knowledge of speed of convergence, and may well proceed for some time before it is apparent that either we still face divergence, or need more alternative models to increase this convergence rate. But of course, by increasing the number of models we not only increase probability of convergence, we also increase our computational time too.

To illustrate, we shall compare approaches using the same series that Cantarelis and Johnston apply their multi-process method to - the widely-known series from Box and Jenkins [3] relating to concentration readings from a chemical process. The full data set is Series 4, given in the Appendix.

Box and Jenkins fit an ARIMA(0,1,1) process to this series, and estimate the autoregressive parameter ϕ to be 0.7, giving the equivalent form of a steady model with noise variances $V = 0.071$ and $W = 0.0091$, so that $r = 0.128$ (note that this does, coincidentally, fall well inside the boundaries of the Cantarelis and Johnston range for $\{r^{(k)}\}$, ensuring reasonable convergence properties of their multi-process approach). We choose to fit several starting choices of V and W, with V chosen from the range (0.01,0.1,1) and W from (0.0015,0.015,0.15). Hence r takes values from (0.0015,0.015,0.15,1.5,15), thus varying by factors of 100, 10, 0.1 and 0.01 from approximately its *a priori* 'true' value (according to the literature that has so far analysed it, that is!). m_0 is taken as 17.0 in each model, and to avoid biassing the convergence rates of either approach we

further take the relatively 'uninformative' prior of $C_0 = 10$ (corresponding to the Cantarelis and Johnston guideline of $100V_0$ for $V_0 = 0.1$). Comparisons are made by calculating the feedback estimates of V and W every 10 time points, using equations 4.14 and 4.15, and sample one step-ahead forecast error variance $S_1^2 = S^2(\hat{e}_t(1|D_t, V, W)) = \frac{1}{n-1}\sum_{i=1}^{n}((Y_i - f_i) - \bar{\hat{e}}_1)^2$ for $\bar{\hat{e}}_1 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_i)$, and covariance $\widehat{C}_{12} = \frac{1}{n-1}\sum_{i=2}^{n}(Y_{i-1} - f_{i-1})(Y_{i-2} - f_{i,2})$. The full results, including Cantarelis and Johnston's analysis for comparison, are given in Table (4.3):

Table 4.3: Results of fitting various steady models to chemical process readings, Series 4 (see Appendix).

				Time t							
	V	W	Ests.	10	20	30	40	80	120	160	197
Analysis using equations 4.14 and 4.15, together with S_1^2 and $\widehat{C}_{12}$	0.01	0.0015	$\widehat{V}_t$	.076	.073	.061	.062	.099	.080	.066	.060
			$\widehat{W}_t$	.0469	.0241	.0212	.0180	.0078	.0102	.0118	.0185
		0.015	$\widehat{V}_t$	.066	.070	.059	.060	.098	.080	.065	.059
			$\widehat{W}_t$	.0722	.0283	.0264	.0218	.0097	.0114	.0128	.0195
		0.15	$\widehat{V}_t$	.059	.066	.057	.058	.097	.079	.065	.058
			$\widehat{W}_t$	.0824	.0399	.0342	.0274	.0124	.0132	.0142	.0210
	0.1	0.0015	$\widehat{V}_t$	.081	.068	.058	.060	.099	.081	.066	.059
			$\widehat{W}_t$	.0226	.0324	.0258	.0212	.0092	.0111	.0126	.0192
		0.015	$\widehat{V}_t$	.076	.073	.061	.062	.099	.081	.066	.060
			$\widehat{W}_t$	.0469	.0252	.0219	.0185	.0080	.0103	.0119	.0186
		0.15	$\widehat{V}_t$	.066	.070	.060	.061	.098	.080	.065	.059
			$\widehat{W}_t$	.0722	.0302	.0275	.0226	.0101	.0116	.0130	.0197
	1	0.0015	$\widehat{V}_t$	.083	.066	.057	.059	.099	.081	.066	.059
			$\widehat{W}_t$	.0081	.0283	.0269	.0221	.0096	.0113	.0128	.0194
		0.015	$\widehat{V}_t$	.082	.070	.059	.061	.099	.081	.066	.060
			$\widehat{W}_t$	.0223	.0327	.0260	.0213	.0093	.0111	.0126	.0193
		0.15	$\widehat{V}_t$	.077	.074	.062	.062	.099	.081	.066	.060
			$\widehat{W}_t$	.0467	.0287	.0240	.0200	.0088	.0108	.0123	.0190
Cantarelis and Johnston analysis			C_t	.36	.26	.24	.22	.24	.21	.20	.21
			$\widehat{Q}_t$	.161	.119	.104	.100	.127	.109	.097	.100
			$\widehat{V}_t$	.121	.088	.074	.071	.100	.083	.069	.071
			$\widehat{W}_t$	.0039	.0046	.0056	.0059	.0039	.0056	.0070	.0077

There are several notable conclusions to be drawn here. Firstly, the speed of convergence of our exact feedback equations is extremely rapid, irrespective of how ill-informed our initial estimates of V and W are. Even after only 10 data points have been received, the range of $\hat{V}_{10}$ is merely (0.059, 0.083) - coming from the grossest initial r estimates of 15 and 0.0015 respectively - and for $\widehat{W}_{10}$ the range is (0.0081, 0.0824), again from the most inaccurate initial estimates of r. All of the $\hat{V}_{10}$ estimates are in extremely close agreement with their respective final $\hat{V}_{197}$ values, and even those for the (more sensitive) $\widehat{W}_{10}$ differ from their respective $\widehat{W}_{197}$ estimates by not more than a factor of 4; moving to the next time point this factor is inside 2. The corresponding convergences for Cantarelis and Johnston's analysis is noticeably poorer for $\hat{V}$, and is no better at time $t = 20$ than even our *worst* misspecification feedback at this point.

The second, easily overlooked, point to remember here is that the Cantarelis and Johnston analysis is not only taking significantly more effort and computer time throughout, but the multi-process approach also means that the feedback estimates they are producing are, at each stage, their *best* estimates, calculated from a probability weighting of *all* the available information at that time. This is in extremely stark contrast to our exact approach, which yields remarkably accurate (relative to the final values) feedback estimates $\hat{V}$ and $\widehat{W}$ immediately and *for each individual model.* Indeed, were we to perform a similar probability weighting analysis with just any two differing initial r specifications from those above, we would receive even more accurate feedbacks throughout the analysis.

Finally, to return to an aside in the previous paragraph, the final estimates require some interpretation. Several features must be observed here: the most

obvious is undoubtedly that the two methods have produced quite different $\hat{r}_{197}$ values. Whereas the various $\widehat{V}_{197}$ estimates are promisingly close, Cantarelis and Johnston's $\widehat{W}_{197}$ estimate of 0.0077 corresponds to $\hat{r}_{197} = 0.108$, compared to Box-Jenkins' $r = 0.128$, and (assuming $\widehat{V}_{197} = 0.06$ and $\widehat{W}_{197} = 0.019$) our value of $\hat{r}_{197} = 0.317$. Further, the patterns in the feedback estimates are reassuringly similar, especially so throughout our exact feedback calculations, but also across both approaches, with a noticeable 'blip' around $t = 80$, where all the $\hat{r}$ estimates decrease dramatically due to a large increase in $\widehat{V}$ to about 0.1, and a decrease in $\widehat{W}$ of around twofold. There is also something of a discernible change around the end of the series, this time due to $\widehat{W}$ apparently increasing, and it does appear that the small bias evident in table 4.2 is visible here too; perhaps the non-linearity of equation 4.15 is creating a slight inflation in the estimates $\{\widehat{W}\}$.

There are few conclusions that we can draw with confidence, therefore - firstly, both approaches appear to be equally sensitive to changes in the series (there is a relatively large 'wobble' in the data through $t = 63, 64, 65$, and again through $t = 190, 191, 192, 193$), and, secondly, the 'true' value V_0 is surely quite close to 0.07 after all (without trying to place a confidence interval on this opinion!). As for $\widehat{W}$, even allowing for a certain level of bias, the quite remarkable similarity after $t = 120$ in both value and behaviour of this parameter, across *all* the models using the exact analysis, lends huge weight to the opinion that the 'true' W_{197} value is rather larger than certainly the 0.0077 obtained by Cantarelis and Johnston, and even the 0.0091 of Box-Jenkins. There is further evidence to support this when we see how the former's estimate of

W has risen steadily after $t = 120$. One would like to believe that our exact algebraic analyses have all converged to the most accurate estimate possible of $W_{197} \simeq 0.019$, but it must be remembered that there is no issue of "who is right", of course, since we are fitting nothing more than an approximate and hugely over-simplistic mathematical model to a complicated physical system, and one which is evidently dynamic and constantly changing.

There is one final and undeniable conclusion, however: not only have we avoided the complexity of the Cantarelis and Johnston approach, and finally shown that there are exact relations to be found in the steady model for estimating the variances therein, but we have *guaranteed* rapid convergence to these estimates for *whatever* prior specifications we make, by finally avoiding the issue of sub-optimal filtering.

This entire method of on-line feedback estimation of V and W stemmed from equation 4.8 in Theorem 4.3. It is evidently applicable to all constant TSDLMs, however, and so we can extend the ideas from the steady model analysis to higher-dimensional models.

4.4 Example 2: the 2-dimensional TSDLM

4.4.1 Canonically equivalent forms within the 2-dimensional model

The scalar constant model of the previous section had a forecast function of the form $f_t(k) = a_{t0}$ $(= m_t)$, a constant. By extending the model definition to the particular 2-dimensional form of

$$M = \left\{(1,0)^T, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, V, \mathbf{W}\right\}$$

the forecast function becomes $f_t(k) = a_{t0} + a_{t1}k$, a linear function in $k \geq 0$. This 2-dimensional model special case is known as a 2nd-order polynomial model; note that it is already in canonical form, with a repeated eigenvalue of 1. Writing the updating equations under M as

$$\begin{aligned} Y_t &= \mu_t + v_t\,, \qquad v_t \sim \mathcal{N}(0, V) \\ \mu_t &= \mu_{t-1} + \beta_t + w_{t1} \\ \beta_t &= \beta_{t-1} + w_{t2}\,, \qquad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{W}) \end{aligned}$$

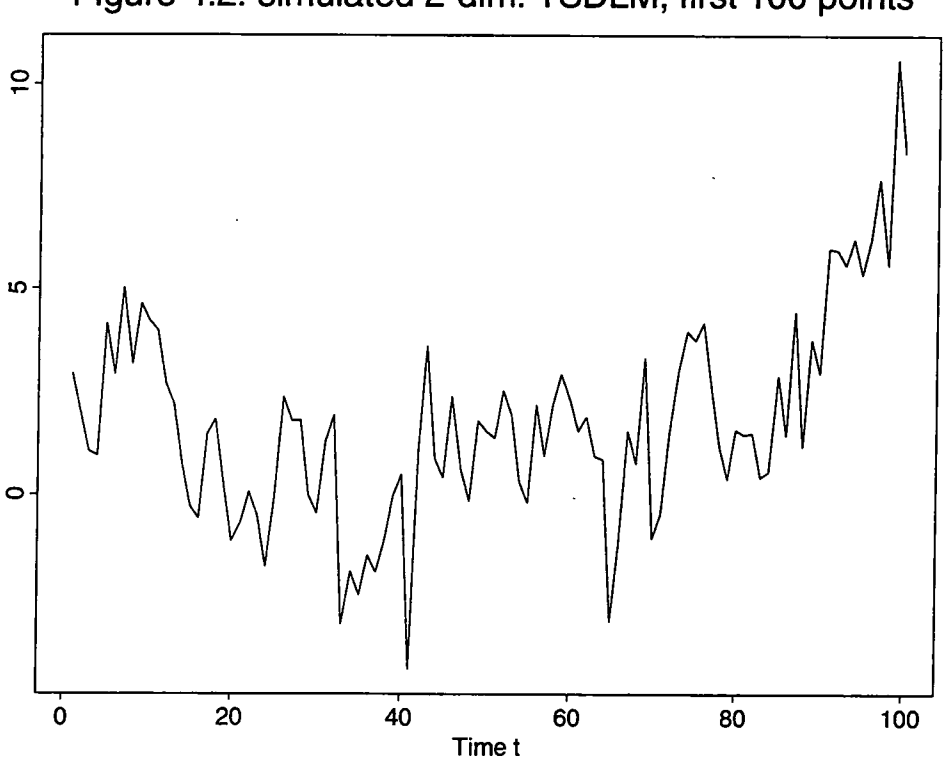
shows that it is interpretable as a linear growth model, widely useful in data evolution where there is an underlying linear trend.

The model is a special case of the more general canonical form for the 2nd-order TSDLM,

$$M_2 = \left\{\mathbf{E} = (1,1)^T, \mathbf{J}_2 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\right\},$$

where now the system evolution matrix has 2 distinct eigenvalues λ_1 and λ_2. We shall consider this general model - much of the following calculation is directly applicable to the 2nd-order polynomial model above, although it must be remembered that it is not simply a case of substituting $\lambda_1 = 1,\ \lambda_2 = 1$, for additionally we now have $\mathbf{E} = (1,1)^T \neq \mathbf{E}_2 = (1,0)^T$, with $\mathbf{J}_2$ no longer diagonal. Hence the precise values within the results are altered, but the principles applied during their calculation are equally applicable to any model in this class.

Figure 4.2: simulated 2-dim. TSDLM, first 100 points



We consider the above form for M_2 with eigenvalues λ_1 and λ_2 of $\mathbf{J}$ being ≤ 1 (else the data evolution explodes exponentially). Figure 4.2 shows the first 100 data points of a simulated series with $\lambda_1 = 1,\ \lambda_2 = 0.5,\ V = V_0 = 1,\ \mathbf{W} =$

$\mathbf{W}_0 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$ (and $\mathbf{m}_0 = (1,1)^T$), which is the particular model M_0 we use throughout this example (Series 5, listed in the Appendix).

Returning to the more general form M_2, then from Lemma 3.2, $\mathbf{E}^T\delta\mathbf{C} = \mathbf{0}$ implies $(\delta C_{11} + \delta C_{12},\ \delta C_{12} + \delta C_{22}) = \mathbf{0}$, and so the perturbation in $\mathbf{C}$ is

$$\delta\mathbf{C} = \begin{pmatrix} \delta C_{22} & -\delta C_{22} \\ -\delta C_{22} & \delta C_{22} \end{pmatrix},$$

giving the same forecast distribution. Hence calculating $\delta\mathbf{W}$ from Lemma 3.1 gives

$$\mathbf{J}_2\delta\mathbf{C}\mathbf{J}_2^T = \begin{pmatrix} \lambda_1^2\delta C_{22} & -\lambda_1\lambda_2\delta C_{22} \\ -\lambda_1\lambda_2\delta C_{22} & \lambda_2^2\delta C_{22} \end{pmatrix}$$

$$\Rightarrow \delta\mathbf{W} = \delta\mathbf{C} - \mathbf{J}_2\delta\mathbf{C}\mathbf{J}_2^T = \begin{pmatrix} (1-\lambda_1^2)\delta C_{22} & (\lambda_1\lambda_2 - 1)\delta C_{22} \\ (\lambda_1\lambda_2 - 1)\delta C_{22} & (1-\lambda_2^2)\delta C_{22} \end{pmatrix}$$

and so again, $\mathbf{W}$ is only fixed in 2 degrees of freedom, as we can take any two model forms

$$\{\mathbf{E}, \mathbf{J}_2, V, \mathbf{W}\} \text{ and } \{\mathbf{E}, \mathbf{J}_2, V, \mathbf{W}' = \mathbf{W} + \delta\mathbf{W}\}$$

for

$$\delta\mathbf{W} = \begin{pmatrix} (1-\lambda_1^2)aW_2 & (\lambda_1\lambda_2 - 1)aW_2 \\ (\lambda_1\lambda_2 - 1)aW_2 & (1-\lambda_2^2)aW_2 \end{pmatrix},$$

and their forecast distributions will be identical. Hence we choose $\mathbf{W} = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}$ as our most convenient representation for the system evolution error variance

matrix, further requiring the constant a to satisfy (so that $\mathbf{W} + \delta\mathbf{W}$ remains semi-positive definite)

$$a^2 W_2(\lambda_1 - \lambda_2)^2 - a(W_1(1-\lambda_2^2) + W_2(1-\lambda_1^2)) - W_1 \leq 0 \ .$$

Note that this diagonal representation of $\mathbf{W}$ will always be feasible in our particular choice of model when, taking the original variance matrix $\mathbf{W}' = \begin{pmatrix} W_1' & W_3' \\ W_3' & W_2' \end{pmatrix}$, we have $W_3' \geq -\frac{2}{3}W_2'$. For then with $\lambda_1 = 1,\ \lambda_2 = 0.5$, the general form for $\delta\mathbf{W}$ becomes equal to $\begin{pmatrix} 0 & -0.5aW_2 \\ -0.5aW_2 & 0.75aW_2 \end{pmatrix}$, so that $aW_2 = -2W_3'$ and hence $0.75aW_2 = -\frac{3}{2}W_3'$. Thus $\mathbf{W} = \begin{pmatrix} W_1' & 0 \\ 0 & W_2' - 0.75aW_2 \end{pmatrix} = \begin{pmatrix} W_1' & 0 \\ 0 & W_2' + \frac{3}{2}W_3' \end{pmatrix}$, which is a valid diagonal form for $\mathbf{W}$ when $W_3' \geq -\frac{2}{3}W_2'$. Since $W_2' \geq 0$, this is evidently true *whenever* the covariance term in $\mathbf{W}'$ is positive.

4.4.2 Effects of model misspecification

We will now follow closely the approach of section 4.3 relating to the steady model, taking a supposed true model of $M_0 = \{\mathbf{E}, \mathbf{J}_2, V_0, \mathbf{W}_0\}$ and our estimated model $M = \{\mathbf{E}, \mathbf{J}_2, V, \mathbf{W}\}$; the main difference being that there are now three unknowns that we wish to solve for, V_0, and W_1, W_2 in $\mathbf{W}_0$. Hence we must consider three distinct equations from the forecast distribution equations 3.5 and 3.6.

Firstly, theorem 4.3 with equations 4.6 and 4.7 give

$$\mathrm{E}[\hat{e}_t(1|D_t,V,\mathbf{W})\hat{e}_t(2|D_t,V,\mathbf{W})] = (\alpha_1-\beta_1)Q_0 + \mathrm{E}\Big[(\mathrm{E}[Y_{t+1}|D_t,V_0,\mathbf{W}_0]-\mathrm{E}[Y_{t+1}|D_t,V,\mathbf{W}]) \cdot (\mathrm{E}[Y_{t+2}|D_t,V_0,\mathbf{W}_0]-\mathrm{E}[Y_{t+2}|D_t,V,\mathbf{W}])\Big] ,$$

with 4.3 yielding the latter half of the RHS of the expression (remembering $\Lambda_0 = 1, \Lambda_1 = \alpha_1$ from 4.2) from

$$\mathrm{E}[Y_{t+1}|D_t,V_0,\mathbf{W}_0] - \mathrm{E}[Y_{t+1}|D_t,V,\mathbf{W}] = \sum_{j=1}^{2}\left(\beta_j - \hat{\beta}_j\left(\frac{1+\sum_{l=1}^{2}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{2}\hat{\beta}_l \mathrm{B}^l}\right)\right) e_{t+1-j} \tag{4.22}$$

and

$$\begin{aligned}\mathrm{E}[Y_{t+2}|D_t,V_0,\mathbf{W}_0] - \mathrm{E}[Y_{t+2}|D_t,V,\mathbf{W}] &= \alpha_1 \sum_{j=1}^{2}\left(\beta_j - \hat{\beta}_j\left(\frac{1+\sum_{l=1}^{2}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{2}\hat{\beta}_l \mathrm{B}^l}\right)\right) e_{t+1-j} \\ &+ \left(\beta_2 - \hat{\beta}_2\left(\frac{1+\sum_{l=1}^{2}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{2}\hat{\beta}_l \mathrm{B}^l}\right)\right) e_t \, . \end{aligned} \tag{4.23}$$

Denoting $\left(\frac{1+\sum_{l=1}^{2}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{2}\hat{\beta}_l \mathrm{B}^l}\right)$ by Ξ gives

$$\begin{aligned}&\mathrm{E}\Big[(\mathrm{E}[Y_{t+1}|D_t,V_0,\mathbf{W}_0]-\mathrm{E}[Y_{t+1}|D_t,V,\mathbf{W}])(\mathrm{E}[Y_{t+2}|D_t,V_0,\mathbf{W}_0]-\mathrm{E}[Y_{t+2}|D_t,V,\mathbf{W}])\Big] \\ &= \mathrm{E}\Big[\Big((\beta_1-\hat{\beta}_1\Xi)e_t+(\beta_2-\hat{\beta}_2\Xi)e_{t-1}\Big)\Big(\alpha_1\Big((\beta_1-\hat{\beta}_1\Xi)e_t+(\beta_2+\hat{\beta}_2\Xi)e_{t-1}\Big)+(\beta_2-\hat{\beta}_2\Xi)e_t\Big)\Big] \, .\end{aligned}$$

However, noting that the first term in Ξ will be 1 (the coefficient of B^0, effectively), we have

$$\mathrm{E}[(\beta_1 e_t)(\hat{\beta}_1 \Xi e_t)] = \beta_1\hat{\beta}_1 Q_0$$

(all other terms are independent, mean 0), and also $\mathrm{E}[(\beta_1 e_t)(\hat{\beta}_2 \Xi e_{t-1})] = 0$, etc., so that

$$\begin{aligned}
&\mathrm{E}\left[(\mathrm{E}[Y_{t+1}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+1}|D_t, V, \mathbf{W}])(\mathrm{E}[Y_{t+2}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+2}|D_t, V, \mathbf{W}])\right] \\
&= (\alpha_1\beta_1^2 + \alpha_1\beta_2^2 - 2\alpha_1(\beta_1\hat{\beta}_1 + \beta_2\hat{\beta}_2) + \beta_1\beta_2 - \beta_1\hat{\beta}_2 - \beta_2\hat{\beta}_1)Q_0 \\
&+ (\alpha_1\hat{\beta}_1^2 + \alpha_1\hat{\beta}_2^2 + \hat{\beta}_1\hat{\beta}_2)\mathrm{Var}(\Xi e_t) - (2\alpha_1\hat{\beta}_1\beta_2 + \beta_2\hat{\beta}_2)\mathrm{E}[e_{t-1}\Xi e_t] \\
&+ (2\alpha_1\hat{\beta}_1\hat{\beta}_2 + \hat{\beta}_2^2)\mathrm{E}[(\Xi e_t)(\Xi e_{t-1})] \qquad (4.24)
\end{aligned}$$

(since $\mathrm{Var}(\Xi e_{t-1}) = \mathrm{Var}(\Xi e_t)$). Hence

$$\begin{aligned}
\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, \mathbf{W}) &= \mathrm{E}[\hat{e}_t(1|D_t, V, \mathbf{W})\hat{e}_t(2|D_t, V, \mathbf{W})] \\
&= (\alpha_1 - \beta_1)Q_0 + 4.24\ . \qquad (4.25)
\end{aligned}$$

Next, considering $\mathrm{Var}(\hat{e}_t(1|D_t, V, \mathbf{W}))$ gives, from 4.22 and 4.6 in 4.8,

$$\begin{aligned}
\mathrm{Var}(\hat{e}_t(1|D_t, V, \mathbf{W})) &= \mathrm{Var}(e_t(1|D_t, V_0, \mathbf{W}_0)) + \mathrm{Var}(\mathrm{E}[Y_{t+1}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+1}|D_t, V, \mathbf{W}]) \\
&= Q_0 + \mathrm{Var}\left((\beta_1 - \hat{\beta}_1\Xi)e_t + (\beta_2 - \hat{\beta}_2\Xi)e_{t-1}\right) \\
&= (1 + \beta_1^2 - 2\beta_1\hat{\beta}_1 + \beta_2^2 - 2\beta_2\hat{\beta}_2)Q_0 + (\hat{\beta}_1^2 + \hat{\beta}_2^2)\mathrm{Var}(\Xi e_t) \\
&- 2\hat{\beta}_1\beta_2\mathrm{E}[e_{t-1}\Xi e_t] + 2\hat{\beta}_1\hat{\beta}_2\mathrm{E}[\Xi e_t \Xi e_{t-1}]\ . \qquad (4.26)
\end{aligned}$$

Finally, our third expression comes from considering the variance of the 2 step-ahead forecast error, which we find via equations 4.7 and 4.23 in 4.8:

$$\mathrm{Var}(\hat{e}_t(2|D_t, V, \mathbf{W})) = \mathrm{Var}(e_t(2|D_t, V_0, \mathbf{W}_0)) + \mathrm{Var}\Big(\mathrm{E}[Y_{t+2}|D_t, V_0, \mathbf{W}_0] - \mathrm{E}[Y_{t+2}|D_t, V, \mathbf{W}]\Big)$$

$$= ((\alpha_1 - \beta_1)^2 + 1)Q_0 + \mathrm{Var}\left(\alpha_1\left((\beta_1 - \hat{\beta}_1\Xi)e_t + (\beta_2 - \hat{\beta}_2\Xi)e_{t-1}\right) + (\beta_2 - \hat{\beta}_2\Xi)e_t\right)$$

$$= (1+(\alpha_1-\beta_1)^2+\alpha_1^2(\beta_1^2+\beta_2^2)-2\alpha_1^2(\beta_1\hat{\beta}_1+\beta_2\hat{\beta}_2)+2\alpha_1(\beta_1\beta_2-\beta_1\hat{\beta}_2-\hat{\beta}_1\beta_2)-2\beta_2\hat{\beta}_2)Q_0$$

$$+ (\alpha_1^2\hat{\beta}_1^2 + 2\alpha_1\hat{\beta}_1\hat{\beta}_2 + (1+\alpha_1^2)\hat{\beta}_2^2)\mathrm{Var}(\Xi e_t) - (2\alpha_1^2\hat{\beta}_1\beta_2 + 2\alpha_1\hat{\beta}_2\beta_2)\mathrm{E}[e_{t-1}\Xi e_t]$$

$$+ (2\alpha_1^2\hat{\beta}_1\hat{\beta}_2 + 2\alpha_1\hat{\beta}_2^2)\mathrm{E}[(\Xi e_t)(\Xi e_{t-1})]\ . \qquad (4.27)$$

To evaluate the three expressions 4.25, 4.26 and 4.27 we must expand

$$\Xi = \left(\frac{1+\sum_{l=1}^{2}\beta_l \mathrm{B}^l}{1+\sum_{l=1}^{2}\hat{\beta}_l \mathrm{B}^l}\right) = (1+\beta_1\mathrm{B}+\beta_2\mathrm{B}^2)(1+\hat{\beta}_1\mathrm{B}+\hat{\beta}_2\mathrm{B}^2)^{-1}$$

$$= b_0 + b_1\mathrm{B} + b_2\mathrm{B}^2 + \ldots$$

$$\text{where } b_0 = 1$$

$$b_1 = (\beta_1 - \hat{\beta}_1)$$

$$b_2 = (\hat{\beta}_1^2 - \hat{\beta}_2 - \beta_1\hat{\beta}_1 + \beta_2)$$

$$\text{and, in general } b_k = b'_k + \beta_1 b'_{k-1} + \beta_2 b'_{k-2},\quad k \geq 2, \qquad (4.28)$$

where, in turn,

$$b'_k = (-1)^k\left(\hat{\beta}_1^k - (k-1)\hat{\beta}_1^{k-2}\hat{\beta}_2 + \frac{1}{2}(k-3)(k-2)\hat{\beta}_1^{k-4}\hat{\beta}_2^2 - \ldots\right.$$

$$\left. \ldots + (-1)^i\binom{k-i}{i}\hat{\beta}_1^{k-2i}\hat{\beta}_2^i + \ldots + \begin{cases} (-1)^{\frac{k+1}{2}}(\frac{k+1}{2})\hat{\beta}_1\hat{\beta}_2^{\frac{k-1}{2}}, & k \text{ odd} \\ (-1)^{\frac{k}{2}}\hat{\beta}_2^{\frac{k}{2}}, & k \text{ even}\ . \end{cases}\right) \qquad (4.29)$$

(Note that the b'_k coefficients are entirely known, being in terms of $\hat{\beta}_1, \hat{\beta}_2$ alone.)

Hence

$$\mathrm{E}[e_{t-1}\Xi e_t] = \mathrm{E}[e_{t-1}b_1 e_{t-1}] \quad \text{since all other terms are 0}$$

$$\Rightarrow \mathrm{E}[e_{t-1}\Xi e_t] \;=\; (\beta_1 - \hat{\beta}_1)Q_0 \;; \tag{4.30}$$

$$\mathrm{Var}(\Xi e_t) = \mathrm{E}[(\Xi e_t)^2] \;=\; \mathrm{E}[b_0^2 e_t^2 + b_1^2 e_{t-1}^2 + \ldots]$$

$$\text{(again, all other terms are 0)} \qquad = \left(\sum_{k=0}^{\infty} b_k^2\right) Q_0 \;; \tag{4.31}$$

$$\text{and} \quad \mathrm{E}[(\Xi e_t)(\Xi e_{t-1})] = \left(\sum_{k=1}^{\infty} b_k b_{k-1}\right) Q_0 \;, \tag{4.32}$$

for b_k defined as above in 4.28.

It is interesting to show - although being somewhat lengthy it is left to the reader - that when $\hat{\beta}_1 = \beta_1$, $\hat{\beta}_2 = \beta_2$, we have $b_k = 0$ for $k = 1, 2, \ldots$, and so these expressions for the various terms in 4.25, 4.26 and 4.27 all reduce the bias to zero in each forecast distribution expression when we take $M = M_0$, which leaves the intuitive results in this case of

$$\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, \mathbf{W}) \underset{M_0}{\overset{M}{\rightarrow}} \mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V_0, \mathbf{W}_0)$$

$$\text{and} \quad \mathrm{Var}(\hat{e}_t(k|D_t, V, \mathbf{W})) \underset{M_0}{\overset{M}{\rightarrow}} \mathrm{Var}(e_t(k|D_t, V_0, \mathbf{W}_0)) \;.$$

Now equations 4.28 and 4.29 together give, through substitution into 4.31,

$$\mathrm{Var}(\Xi e_t) = \left\{\sum_{k=0}^{\infty}(b'_k)^2 + \left(2\sum_{k=1}^{\infty} b'_k b'_{k-1}\right)\beta_1 + \left(2\sum_{k=2}^{\infty} b'_k b'_{k-2}\right)\beta_2 \right.$$

$$\left. + \left(2\sum_{k=2}^{\infty} b'_{k-1} b'_{k-2}\right)\beta_1\beta_2 + \left(\sum_{k=1}^{\infty}(b'_{k-1})^2\right)\beta_1^2 + \left(\sum_{k=2}^{\infty}(b'_{k-2})^2\right)\beta_2^2 \right\} Q_0$$

$$= \left\{\sum_{k=0}^{\infty}(b'_k)^2(1+\beta_1^2+\beta_2^2) + \left(2\sum_{k=1}^{\infty} b'_k b'_{k-1}\right)(\beta_1\beta_2+\beta_1) + \left(2\sum_{k=2}^{\infty} b'_k b'_{k-2}\right)\beta_2 \right\} Q_0 \;,$$

with similarly

$$\begin{aligned}
\mathrm{E}[(\Xi e_t)(\Xi e_{t-1})] &= \left(\sum_{k=1}^{\infty} b_k b_{k-1}\right) Q_0 \\
&= \left\{\left(\sum_{k=0}^{\infty}(b'_k)^2\right)(\beta_1+\beta_1\beta_2)+\left(\sum_{k=1}^{\infty} b'_k b'_{k-1}\right)(1+\beta_2+\beta_1^2+\beta_2^2)\right. \\
&+ \left.\left(\sum_{k=2}^{\infty} b'_k b'_{k-2}\right)(\beta_1+\beta_1\beta_2)+\left(\sum_{k=3}^{\infty} b'_k b'_{k-3}\right)\beta_2\right\} Q_0
\end{aligned}$$

Finally, we denote $\sum_{k=j}^{\infty}(b'_k)(b'_{k-j})$ by s_j, for $j = 0, 1, 2, 3$, and then substituting these last two expressions, together with 4.30, into equations 4.25, 4.26 and 4.27 in turn results in

$$\begin{aligned}
\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, \mathbf{W}) &= \left\{\left[\alpha_1 + s_0(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)+s_1(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\right]\right. \\
&+ \left[2s_1(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)-2\alpha_1\hat\beta_1-\hat\beta_2-1+(s_0+s_2)(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\right]\beta_1 \\
&+ \left[2s_2(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)-2\alpha_1\hat\beta_2-\hat\beta_1+2\alpha_1\hat\beta_1^2+\hat\beta_1\hat\beta_2+(s_1+s_3)(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\right]\beta_2 \\
&+ \left[2s_1(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)-2\alpha_1\hat\beta_1-\hat\beta_2-1+(s_0+s_2)(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\right]\beta_1\beta_2 \\
&+ \left[\alpha_1 + s_0(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)+s_1(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\right]\beta_1^2 \\
&+ \left.\left[\alpha_1 + s_0(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)+s_1(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\right]\beta_2^2\right\} Q_0 \,.
\end{aligned}$$

So, further defining the entirely known constants

$$\begin{aligned}
c_1 &\equiv s_0(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)+\alpha_1+s_1(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\,, \\
c_2 &\equiv 2s_1(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)-2\alpha_1\hat\beta_1-\hat\beta_2-1+(s_0+s_2)(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\,, \\
c_3 &\equiv 2s_2(\alpha_1(\hat\beta_1^2+\hat\beta_2^2)+\hat\beta_1\hat\beta_2)-2\alpha_1\hat\beta_2-\hat\beta_1+2\alpha_1\hat\beta_1^2+\hat\beta_1\hat\beta_2+(s_1+s_3)(2\alpha_1\hat\beta_1\hat\beta_2+\hat\beta_2^2)\,,
\end{aligned}$$

we reach the quadratic

$$\mathrm{Cov}(Y_{t+1}, Y_{t+2}|D_t, V, \mathbf{W}) = \left(c_1 + c_2\beta_1 + c_3\beta_2 + c_2\beta_1\beta_2 + c_1\beta_1^2 + c_1\beta_2^2\right) Q_0 \,. \tag{4.33}$$

Similarly, we can also produce

$$\mathrm{Var}(\hat{e}_t(1|D_t, V, \mathbf{W})) = (c_4 + c_5\beta_1 + c_6\beta_2 + c_5\beta_1\beta_2 + c_4\beta_1^2 + c_4\beta_2^2)Q_0 \,, \tag{4.34}$$

$$\begin{aligned} \text{for} \quad c_4 &= s_0(\hat{\beta}_1^2 + \hat{\beta}_2^2) + 1 + 2s_1\hat{\beta}_1\hat{\beta}_2 \,, \\ c_5 &= 2s_1(\hat{\beta}_1^2 + \hat{\beta}_2^2) - 2\hat{\beta}_1 + 2(s_0 + s_2)\hat{\beta}_1\hat{\beta}_2 \\ \text{and} \quad c_6 &= 2s_2(\hat{\beta}_1^2 + \hat{\beta}_2^2) - 2\hat{\beta}_2 + 2\hat{\beta}_1^2 + 2(s_1 + s_3)\hat{\beta}_1\hat{\beta}_2 \,, \end{aligned}$$

and

$$\mathrm{Var}(\hat{e}_t(2|D_t, V, \mathbf{W})) = (c_7 + c_8\beta_1 + c_9\beta_2 + c_8\beta_1\beta_2 + c_7\beta_1^2 + c_7\beta_2^2)Q_0 \,, \tag{4.35}$$

for

$$c_7 = s_0(\alpha_1^2\hat{\beta}_1^2 + 2\alpha_1\hat{\beta}_1\hat{\beta}_2 + (1+\alpha_1^2)\hat{\beta}_2^2) + 1 + \alpha_1^2 + 2s_1(\alpha_1^2\hat{\beta}_1\hat{\beta}_2 + \alpha_1\hat{\beta}_2^2) \,,$$

$$c_8 = 2s_1(\alpha_1^2\hat{\beta}_1^2 + 2\alpha_1\hat{\beta}_1\hat{\beta}_2 + (1+\alpha_1^2)\hat{\beta}_2^2) + 2\alpha_1(1 - \alpha_1\hat{\beta}_1 - \hat{\beta}_2) + (s_0+s_2)(\alpha_1^2\hat{\beta}_1\hat{\beta}_2 + \alpha_1\hat{\beta}_2^2) \,,$$

$$c_9 = 2s_2(\alpha_1^2\hat{\beta}_1^2 + 2\alpha_1\hat{\beta}_1\hat{\beta}_2 + (1+\alpha_1^2)\hat{\beta}_2^2) + 2\alpha_1\hat{\beta}_1(\hat{\beta}_2 - 1) + 2\alpha_1^2(\hat{\beta}_1^2 - \hat{\beta}_2) - 2\hat{\beta}_2$$
$$+ 2(s_1 + s_3)(\alpha_1^2\hat{\beta}_1\hat{\beta}_2 + \alpha_1\hat{\beta}_2^2) \,.$$

4.4.3 Solving for V_0, W_1 and W_2

Equations 4.33, 4.34 and 4.35 are three simultaneous quadratic equations in the only unknown elements present - Q_0, β_1 and β_2 - since the constants c_i, $i = 1, \ldots, 9$, are wholly calculable directly from the b'_k coefficients, which are in turn functions of the known $\hat{\beta}_1$ and $\hat{\beta}_2$ only.

We proceed by calculating sample estimates S_1^2, S_2^2 and $\hat{C}_{12}$ of the one and two step-ahead forecast errors, and the lag-one covariance, respectively. Given these sample estimates, we can solve numerically for the unknowns Q_0, β_1 and β_2 from our set of simultaneous quadratic equations. Then further, we know that under the true model $M_0 = \{\mathbf{E}, \mathbf{J}_2, V_0, \mathbf{W}_0\}$ the eigenvalues of $\mathbf{J}_2$ are λ_1 and λ_2, and those of $\mathbf{H}_0 = (\mathbf{I}_2 - \mathbf{A}_0\mathbf{E}^T)\mathbf{J}_2 = \begin{pmatrix} (1-A_1)\lambda_1 & -A_1\lambda_2 \\ -A_2\lambda_1 & (1-A_2)\lambda_2 \end{pmatrix}$ are ρ_1 and ρ_2, where the convergent form of the adaptive coefficient $\mathbf{A}_0$ is denoted by $(A_1, A_2)^T$. Therefore

$$\begin{aligned} \beta_1 = -(\rho_1 + \rho_2) &= -\text{trace}(\mathbf{H}_0) \\ &= -(\lambda_1 + \lambda_2 - A_1\lambda_1 - A_2\lambda_2) \\ \text{and } \beta_2 = \rho_1\rho_2 &= \det(\mathbf{H}_0) \\ &= \lambda_1\lambda_2(1 - A_1 - A_2) \ . \end{aligned}$$

Solving for $\mathbf{A}_0$ yields

$$\begin{aligned} A_1 &= \frac{\lambda_1^2 + \lambda_1\beta_1 + \beta_2}{\lambda_1(\lambda_1 - \lambda_2)} \\ \text{and } A_2 &= \frac{\lambda_2^2 + \lambda_2\beta_1 + \beta_2}{\lambda_2(\lambda_2 - \lambda_1)} \ . \end{aligned}$$

Then

$$\begin{aligned}
\mathbf{E}^T\mathbf{R}_0\mathbf{E} = \mathbf{E}^T\mathbf{A}_0Q_0 &= (A_1 + A_2)Q_0 \\
&= \left(\frac{\lambda_1^2\lambda_2 + \lambda_1\lambda_2\beta_1 + \lambda_2\beta_2 - \lambda_1\lambda_2^2 - \lambda_1\lambda_2\beta_1 - \lambda_1\beta_2}{\lambda_1\lambda_2(\lambda_1 - \lambda_2)}\right) Q_0 \\
&= \left(\frac{\lambda_1\lambda_2(\lambda_1 - \lambda_2) + \beta_2(\lambda_2 - \lambda_1)}{\lambda_1\lambda_2(\lambda_1 - \lambda_2)}\right) Q_0 \\
&= \left(1 - \frac{\beta_2}{\lambda_1\lambda_2}\right) Q_0 \,,
\end{aligned}$$

and so

$$V_0 = Q_0 - \mathbf{E}^T\mathbf{R}_0\mathbf{E} = \frac{Q_0\beta_2}{\lambda_1\lambda_2} \,. \tag{4.36}$$

We can calculate $\mathbf{W}_0 = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}$ in a similar fashion; observe that

$$\begin{aligned}
\frac{\mathbf{R}_0\mathbf{E}}{Q_0} &= \mathbf{A}_0 \\
\Rightarrow \begin{pmatrix} R_{11} + R_{12} \\ R_{12} + R_{22} \end{pmatrix} &= \begin{pmatrix} A_1Q_0 \\ A_2Q_0 \end{pmatrix} \quad \left(\text{where } \mathbf{R}_0 = \begin{pmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{pmatrix}\right) \\
\Rightarrow \mathbf{R}_0 &= \begin{pmatrix} A_1Q_0 - R_{12} & R_{12} \\ R_{12} & A_2Q_0 - R_{12} \end{pmatrix} .
\end{aligned}$$

So if $\mathbf{C}_0 = \begin{pmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{pmatrix}$, we have from $\mathbf{C}_0 = \mathbf{R}_0 - \mathbf{A}_0\mathbf{A}_0^TQ_0$ that

$$\begin{aligned}
C_{11} &= A_1Q_0 - R_{12} - A_1^2Q_0 \\
C_{12} &= R_{12} - A_1A_2Q_0 \\
\text{and} \quad C_{22} &= A_2Q_0 - R_{12} - A_2^2Q_0 \,.
\end{aligned}$$

Further, $\mathbf{J}_2\mathbf{C}_0\mathbf{J}_2^T = \begin{pmatrix} \lambda_1^2 C_{11} & \lambda_1\lambda_2 C_{12} \\ \lambda_1\lambda_2 C_{12} & \lambda_2^2 C_{22} \end{pmatrix}$, and so for $\mathbf{W}_0 = \mathbf{R}_0 - \mathbf{J}_2\mathbf{C}_0\mathbf{J}_2^T$ to be diagonal we must have

$$C_{12} = \frac{R_{12}}{\lambda_1\lambda_2} .$$

Substituting this into the three simultaneous equations in C_{11}, C_{12} and C_{22} produces

$$\begin{aligned} C_{11} &= A_1 Q_0 \left(1 - A_1 - \frac{\lambda_1\lambda_2 A_2}{1-\lambda_1\lambda_2}\right) \\ C_{12} &= \frac{A_1 A_2 Q_0}{\lambda_1\lambda_2 - 1} \\ \text{and} \quad C_{22} &= A_2 Q_0 \left(1 - A_2 + \frac{\lambda_1\lambda_2 A_1}{1-\lambda_1\lambda_2}\right) . \end{aligned}$$

Then, ultimately, from

$$\begin{aligned} \mathbf{W}_0\mathbf{E} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} &= \mathbf{R}_0\mathbf{E} - \mathbf{J}_2\mathbf{C}_0\mathbf{J}_2^T\mathbf{E} \\ &= \mathbf{A}_0 Q_0 - \mathbf{J}_2\mathbf{C}_0\mathbf{J}_2^T\mathbf{E} \end{aligned}$$

we have

$$W_1 = A_1 Q_0 \left(\frac{1 + \lambda_1\lambda_2(A_2 - 1) + \lambda_1^2(A_1 - 1) + \lambda_1^3\lambda_2(1 - A_1 - A_2)}{1 - \lambda_1\lambda_2}\right) \quad (4.37)$$

and

$$W_2 = A_2 Q_0 \left(\frac{1 + \lambda_1\lambda_2(A_1 - 1) + \lambda_2^2(A_2 - 1) + \lambda_1\lambda_2^3(1 - A_1 - A_2)}{1 - \lambda_1\lambda_2}\right) , \quad (4.38)$$

$$\text{for} \quad A_1 = \frac{\lambda_1^2 + \lambda_1\beta_1 + \beta_2}{\lambda_1(\lambda_1 - \lambda_2)}$$

$$\text{and} \quad A_2 = \frac{\lambda_2^2 + \lambda_2\beta_1 + \beta_2}{\lambda_2(\lambda_2 - \lambda_1)} .$$

So knowing Q_0, β_1 and β_2 allows us to calculate V_0, W_1 and W_2 all relatively straightforwardly. The accuracy of these variance estimates again depends solely upon the convergence of the sample values S_1^2, S_2^2 and $\hat{C}_{12}$; if these sample estimates were to equal their respective theoretical *biassed* values, then our feedback estimates $\widehat{V}, \widehat{W}_1$ and $\widehat{W}_2$ would all be *exact.*

4.4.4 Simulation results

The model chosen is as at the start of this section, namely

$$M_0 = \left\{ \mathbf{E}, \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, V_0 = 1, \mathbf{W}_0 = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} \right\},$$

and the first 100 data points for a simulated series of length 1000 for these specifications were shown in Figure 4.2. Various model misspecifications were made before forecasting this data set for its last 950 points (the first 50 points of each analysis are ignored to let the model reach its limiting state); all of V_0, W_1 and W_2 were varied both individually and in pairs, with each misspecification coming from factors of either 10 or 100.

The feedback estimates $\widehat{V}$ of V_0, and $\widehat{W}_1$ and $\widehat{W}_2$ of W_1 and W_2, were then calculated via equations 4.36, 4.37 and 4.38, having already obtained the sample variances S_1^2 and S_2^2 together with the sample lag one covariance $\hat{C}_{12}$, and having

solved for the estimates $\widehat{Q}$, $\widehat{\beta_1^0}$ and $\widehat{\beta_2^0}$ of Q_0, β_1 and β_2 from equations 4.33, 4.34 and 4.35.

The results are shown in full in Table 4.4. Note that under the true model M_0, we have $\mathbf{A}_0 = \begin{pmatrix} 0.399 \\ 0.284 \end{pmatrix}$, and so

$$\mathbf{H}_0 = (\mathbf{I}_2 - \mathbf{A}_0\mathbf{E})\mathbf{J}_2 = \begin{pmatrix} 0.601 & -0.199 \\ -0.284 & 0.358 \end{pmatrix},$$

implying that $\rho_1 = 0.747$ and $\rho_2 = 0.213$. Thus

$$\begin{aligned} \beta_1 &= -(\rho_1 + \rho_2) = -0.960 \\ \text{and} \quad \beta_2 &= \rho_1\rho_2 = 0.159\ , \\ \text{with further} \quad Q_0 &= \mathbf{E}^T\mathbf{R}_0\mathbf{E} + V_0 = 3.147\ . \end{aligned}$$

As a final point, note that it is vital to draw a distinction between the $\hat{\beta}_1$, $\hat{\beta}_2$ and Q values arising from the misspecified model itself, and the feedback estimates $\widehat{\beta_1^0}$, $\widehat{\beta_2^0}$ and $\widehat{Q}$ (obtained through solution of 4.33, 4.34 and 4.35) of the true values (as under M_0) β_1, β_2 and Q_0. It is these feedback estimates which are given in the table.

Table 4.4: results for various fitted models M on simulated Series 5 (see Appendix) of length $n = 950$, with $V_0 = 1$, $\mathbf{W}_0 = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$.

V	W_1	W_2	$\widehat{C}_{12}$	S_1^2	S_2^2	$\widehat{Q}$	$\widehat{\beta}_1^0$	$\widehat{\beta}_2^0$	$\sum_{i=1}^3 f_i^2$	$\widehat{V}$	$\widehat{W}_1$	$\widehat{W}_2$
						obtained from				from	from	from
						4.33, 4.34 and 4.35				4.36	4.37	4.38
1	0.5	1	1.716	3.132	4.078	3.133	-0.952	0.159	1.14×10^{-5}	1.00	0.54	0.94
	5		2.126	3.448	4.608	3.142	-0.948	0.129	1.75×10^{-5}	0.81	0.41	1.34
	50		2.477	3.746	5.024	3.125	-0.948	0.130	1.93×10^{-5}	0.82	0.42	1.32
	0.05		1.953	3.351	4.331	3.112	-0.958	0.111	9.42×10^{-6}	0.69	0.29	1.61
	0.005		3.065	4.253	5.704	3.112	-0.957	0.091	1.82×10^{-5}	0.57	0.22	1.86
1	0.5	10	1.754	3.166	4.143	3.114	-0.960	0.112	1.27×10^{-5}	0.70	0.29	1.61
		100	2.383	3.586	5.112	3.119	-0.961	0.096	1.82×10^{-5}	0.60	0.23	1.82
		0.1	1.783	3.191	4.193	3.158	-0.946	0.145	1.10×10^{-5}	0.91	0.50	1.11
		0.01	1.796	3.204	4.204	3.158	-0.946	0.142	1.09×10^{-5}	0.90	0.49	1.15
0.1	0.5	1	1.822	3.243	4.180	3.147	-0.943	0.156	1.42×10^{-5}	0.98	0.57	0.93
0.01			1.857	3.284	4.210	3.148	-0.943	0.155	1.48×10^{-5}	0.97	0.57	0.95
10			1.994	3.526	4.262	3.112	-0.961	0.132	7.82×10^{-6}	0.82	0.36	1.36
100			3.234	4.916	5.370	3.095	-0.955	0.091	2.35×10^{-3}	0.56	0.23	1.85
10	0.5	0.1	2.029	3.597	4.278	3.114	-0.962	0.137	6.57×10^{-6}	0.85	0.38	1.30
1	5	10	1.822	3.243	4.180	3.147	-0.943	0.156	1.42×10^{-5}	0.98	0.57	0.93

4.4.5 Conclusions

There are a number of things to note in relation to Table 4.4.

(i) The first line in the table (for correct model specification of $V = V_0$ and $\mathbf{W} = \mathbf{W}_0$) has produced remarkably precise feedback estimates $\widehat{V}$ and $\widehat{\mathbf{W}}$, and so seems to suggest the methodology is accurate, at least!

(ii) The tenth and fifteenth lines, for $V = 0.1$, $\mathbf{W} = \mathbf{W}_0 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$ and $V = V_0 = 1$, $\mathbf{W} = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}$ respectively, illustrate through their parity how it is again the *ratio* of the elements of V and $\mathbf{W}$ that influences the behaviour of the forecast distribution.

(iii) There appears to be evidence of correlation between the estimates $\widehat{W}_1$ and $\widehat{W}_2$; as $\widehat{W}_2$ is inflated above its true value of 1, so $\widehat{W}_1$ consistently underes-

timates its true value of 0.5 (and vice-versa). Moreover there is some evidence, once more, of bias in our variance estimates, most noticeable in $\widehat{W}_2$; this is probably due, again, to the non-linearity throughout equations 4.33 to 4.38.

(iv) The three simultaneous quadratic equations 4.33, 4.34 and 4.35 are rather sensitive to perturbations in the estimates $\widehat{C}_{12}$, S_1^2 and S_2^2. Accordingly, exact solutions for $\widehat{Q}$, $\widehat{\beta_1^0}$ and $\widehat{\beta_2^0}$ were not generally obtainable (producing complex roots), but instead a numerical minimisation method was employed, where the sum of the squares

$$\sum_{i=1}^{3}(f_i(Q_0, \beta_1, \beta_2) - E_i)^2$$

of each quadratic function $f_i(Q_0, \beta_1, \beta_2) - E_i = 0$, for $i = 1, 2, 3$ and $E_1 = S_1^2$, $E_2 = S_2^2$, $E_3 = \widehat{C}_{12}$, is minimised. There are several local minima of $\sum_{i=1}^{3} f_i(Q_0, \beta_1, \beta_2)$, but convergence was generally rapid to the given minimum.

(v) This previous point is, of course, a fundamental issue with the success of the approach. When minimising $\sum_{i=1}^{3}(f_i(Q_0, \beta_1, \beta_2) - E_i)^2$, we must feed in some starting values for each of Q_0, β_1 and β_2, and it is logical to presume that the practitioner will make his initial guess equal to his current estimation of these parameters - namely $\{Q, \hat{\beta}_1, \hat{\beta}_2\}$ - arising from his *initial* model specification; a practitioner will not alter his initial beliefs in the variances within the model unless he has good cause to, which is, evidently, the aim of this entire methodology! Thus the values given in the table for $\widehat{Q}$, $\widehat{\beta_1^0}$ and $\widehat{\beta_2^0}$ are, in fact, the estimates that are obtained from minimisation of $\sum_{i=1}^{3}(f_i(Q_0, \beta_1, \beta_2) - E_i)^2$ from this starting value of $\{Q, \hat{\beta}_1, \hat{\beta}_2\}$ calculated directly from each misspecified model M.

(vi) Having said this, there is some qualification of this last remark to be made. Occasionally the starting value of $\{Q, \hat{\beta}_1, \hat{\beta}_2\}$ converged to a different local minimum, but all of these local minima went on to produce infeasible (i.e. negative) values of $\hat{V}$ or $\widehat{\mathbf{W}}$ from equations 4.36, 4.37 and 4.38. It does appear, therefore, that our method is robust in so much as it leads to the 'correct' solution by a process of elimination...

...(vii) And equations 4.33, 4.34 and 4.35 have, after one takes heed of the previous remarks, indeed produced reassuringly accurate estimates $\hat{Q}$, $\widehat{\beta_1^0}$ and $\widehat{\beta_2^0}$. Hence the resulting discrepancies in the feedback estimates $\hat{V}$ and $\widehat{\mathbf{W}}$ are evidently down to the sensitivity of equations 4.36, 4.37 and 4.38 to numerical perturbations in these estimates $\hat{Q}$, $\widehat{\beta_1^0}$ and $\widehat{\beta_2^0}$. This is merely further justification of the previous remark, and if anything is a blessing in disguise, for it is this very sensitivity which (seemingly) guarantees only one of the local minima (or the global minimum) will produce feasible variance estimates.

(viii) Penultimately, we should observe that the final on-line variance estimates are all quite agreeable. Some are evidently more so than others, but it must be noted that the least accurate lines in Table 4.4 are again those involving the worst model misspecifications, such as in the fifth and seventh lines, where W_1 and W_2 are out by factors of 100 respectively. Even here the resulting feedback variance estimates are so far removed from the initial variance specifications that the practitioner would be rapidly moved to far more precise specifications, and whence to even more precise estimates $\hat{V}$ and $\widehat{\mathbf{W}}$, and so on.

(ix) 'And so on' leads us to the final simulation. In practice, the practitioner attempting to use this method will not have 950 data points to use at his leisure

whilst waiting for $\widehat{V}$ and $\widehat{\mathbf{W}}$ to converge. Even if he has, he will want confirmation - or otherwise - that his initial variance specifications are reasonable. Hence his adopted practice will be to wait for, say, T time points into the analysis until reasonable convergence of S_1^2, S_2^2 and $\widehat{C}_{12}$ has been reached, before obtaining initial feedback estimates $\widehat{V}_T$ and $\widehat{\mathbf{W}}_T$ up to that time T, which in turn would be used in forecasting the next T time points to obtain second estimates $\widehat{V}_{2T}$ and $\widehat{\mathbf{W}}_{2T}$, 'and so on'. The length of time T is crucial here - and unfortunately it is not particularly easy to obtain the same convergence properties results of the previous section (with respect to the steady model) in the 2-dimensional case. Hence this 'waiting time' T must be estimated by monitoring the sample variances S_1^2 and S_2^2, and covariance $\widehat{C}_{12}$, at each time point, and deciding when 'reasonable' convergence has indeed been reached.

The following Table, 4.5, breaks down the previous analysis for two particular models into time intervals of length 50 and 100, and shows the feedback estimates $\widehat{V}_t$ and $\widehat{\mathbf{W}}_t$ that would be obtained through calculating S_1^2, S_2^2 and $\widehat{C}_{12}$ for the entire preceding analysis up to time t. In both examples, the analysis is terminated when the convergence of the sample variances and lag one covariance has become sufficiently accurate to render all future estimates $\{\widehat{V}_t\}$ - and $\{\widehat{\mathbf{W}}_t\}$ - very similar.

Table 4.5: results for various fitted models M on simulated Series 5, with $V_0 = 1$, $\mathbf{W}_0 = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$, up to different time points t.

Time t	V	W_1	W_2	$\widehat{C}_{12}$	S_1^2	S_2^2	$\widehat{Q}$	$\widehat{\beta}_1^0$	$\widehat{\beta}_2^0$	$\widehat{V}$	$\widehat{W}_1$	$\widehat{W}_2$
							obtained from 4.33, 4.34 and 4.35			from 4.36	from 4.37	from 4.38
50	1	0.5	1	1.361	2.621	3.229	2.592	-0.991	0.193	1.00	0.42	0.60
100				1.372	2.711	3.350	2.690	-1.001	0.175	0.94	0.33	0.89
150				1.803	3.339	4.218	3.286	-0.974	0.221	1.46	0.81	0.23
200				1.636	3.281	4.109	3.256	-1.008	0.153	1.00	0.27	1.41
250				1.652	3.331	4.166	3.331	-1.003	0.190	1.27	0.47	0.89
350				1.709	3.367	4.239	3.346	-0.996	0.231	1.54	0.74	0.26
450				1.664	3.219	4.081	3.213	-0.985	0.158	1.02	0.39	1.19
⋮												
1000				1.716	3.132	4.078	3.133	-0.952	0.159	1.00	0.54	0.94
50	10	0.5	0.1	1.841	2.980	3.560	2.365	-0.964	0.206	0.97	0.55	0.29
100				1.841	3.050	3.603	2.429	-0.984	0.228	1.11	0.57	0.16
150				2.317	3.937	4.731	3.235	-0.995	0.229	1.48	0.71	0.27
200				2.002	3.723	4.397	3.250	-1.011	0.207	1.35	0.50	0.69
250				1.904	3.698	4.336	3.292	-1.028	0.214	1.41	0.46	0.71
350				1.931	3.722	4.362	3.292	-1.028	0.220	1.45	0.49	0.63
450				1.977	3.646	4.306	3.198	-0.992	0.174	1.11	0.42	1.02
⋮												
1000				1.994	3.526	4.262	3.112	-0.961	0.132	0.82	0.36	1.36

The first example was chosen as the true model M_0 to allow us to concentrate on the behaviour of the sample estimates S_1^2, S_2^2 and $\widehat{C}_{12}$ alone. We see that their convergence is relatively slow - even though each time point's estimates $\widehat{V}$, $\widehat{W}_1$ and $\widehat{W}_2$ are still reasonably accurate - with a noticeable blip around $t = 150$ which inflates the step-ahead variances and covariances significantly, thus affecting the on-line estimates $\widehat{V}$ and $\widehat{\mathbf{W}}$ around this time. There is another such feature between $t = 250$ and 350 (but with slightly less effect, as it is further into the analysis), and both these blips are similarly in evidence - with the same inflating effect on the forecast distribution equations - in the second example.

These two perturbations in the data set are apt reminder that even though we are dealing with a simulated series, it is still prone to the kind of irregularities that occur in genuine data series.

4.5 Final discussions

Unlike the 1st-order steady model illustration of the application of Theorem 4.3, we evidently have much slower convergence of the sample variances and covariances when dealing with the 2nd-order TSDLM. There is a further complication in this latter case due to the difficulty of solving the simultaneous quadratic equations 4.33, 4.34 and 4.35, and the sensitivity of the resulting calculations of $\widehat{V}$ and $\widehat{\mathbf{W}}$ from equations 4.36, 4.37 and 4.38. The overall conclusion must be that in the 2-dimensional model, we need a series of length around $n = 100$ before the stability of the sample variances S_1^2 and S_2^2, and the lag-one covariance $\widehat{C}_{12}$, can be relied upon for genuinely accurate feedback variance estimates (and we should remark that even in the more severe model misspecifications, our methodology is producing (relatively) exactly that). The convergence of the sample forecast distribution would undoubtedly become more of an issue when dealing with higher order TSDLMs, as would the method of solution of the set of simultaneous equations required for the complete estimation of V and a diagonal $\mathbf{W}$ (in general we would have a set of $p+1$ equations, each a p^{th}-order polynomial in the β_i terms, when dealing with a p-dimensional model).

However, the overwhelming success of the more-than-useful steady model application, and the still-notable success of the general 2-dimensional TSDLM

example, once an acceptance of the calculational complexity and intricacies has been made, leaves the exciting feeling that in the constant TSDLM, the problem of variance estimation has finally been overcome.

Bibliography

[1] AHN, S.K. (1988). "Distribution for residual autocovariances in multivariate autoregressive models with structured parametrization", *Biometrika* **75**, 590-3.

[2] AMEEN, J.R.M. & HARRISON, P.J. (1985). "Normal Discount Bayesian Models", in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.), North-Holland, Amsterdam, and Valencia University Press.

[3] BOX, G.E.P. & JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

[4] BOX, G.E.P. & PIERCE, D.A. (1970). "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models", *Journal of the American Statistical Association* **65**, 1509-26.

[5] BROADBENT, S. (1979). "One way T.V. advertisements work", *Journal of the Market Research Society* **21**, 139-166.

[6] BROWN, R.G. (1963). *Smoothing, Forecasting and Prediction*, Englewood Cliffs: Prentice-Hall.

[7] CANTARELIS, N. & JOHNSTON, F.R. (1983). "On-Line Variance Estimation for the Steady State Bayesian Forecasting Model", *J. Time Ser. Anal.* **3**, 225-234.

[8] CHATFIELD, C. & PROTHERO, D.L. (1973). "Box-Jenkins Seasonal Forecasting: Problems in a Case-Study" (with discussion), *J. Roy. Statist. Soc.* (Ser. A) **136**, 295-352.

[9] DAVIES, N., TRIGGS, C.M. & NEWBOLD, P. (1977). "Significance levels of the Box-Pierce portmanteau statistic in finite samples", *Biometrika* **64**, 517-22.

[10] DAVIES, N. & NEWBOLD, P. (1979). "Some power studies of a portmanteau test of time series model specification", *Biometrika* **66**, 153-5.

[11] DURBIN, J. (1970). "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables", *Econometrica* **38**, 410-21.

[12] FAMA, E.F & SCHWERT, G.W. (1977). "Asset returns and inflation", *Journal of Financial Economics* **5**, 115-46.

[13] GODBOLE, S.S. (1974). "Kalman Filtering with no *a priori* Information about Noise - White Noise Case: Identification of Covariances", *IEEE Trans. Automat. Contr.* **AC-19**, 561-563.

[14] GODFREY, L.G. (1979). "Testing the adequacy of a time series model", *Biometrika* **66**, 67-72.

[15] GODOLPHIN, E.J. (1980). "A method for testing the order of an autoregressive - moving average process", *Biometrika* **67**, 699-703.

[16] GREEN, M. & HARRISON, P.J. (1973). "Fashion Forecasting for a Mail Order Company using a Bayesian Approach", *Op. Res. Quart.* **24**, 193-205.

[17] HARRISON, P.J. & AKRAM, M. (1983). "Generalised exponentially weighted regression and parsimonious dynamic linear modelling", *Time Series Analysis* **3**, 19-42.

[18] HARRISON, P.J. & STEVENS, C.F. (1971). "A Bayesian Approach to Short-term Forecasting", *Op. Res. Quart.* **22**, 341-362.

[19] HARRISON, P.J. & STEVENS, C.F. (1976). "Bayesian Forecasting" (with discussion), *J. Roy. Statist. Soc.* (Ser. B) **38**, 205-247.

[20] HOLT, C.C. (1957). "Forecasting seasonals and trends by exponentially weighted moving averages", *ONR Research Memorandum* **52**, Carnegie Institute of Technology, Pittsburgh, Pennsylvania.

[21] HOSKING, J.R.M. (1980). "Lagrange-multiplier Tests of Time-Series Models", *Journal of the Royal Statistical Society*, Series B **42**, 170-181.

[22] KALMAN, R.E. (1960). "A new Approach to Linear Filtering and Prediction Problems", *J. of Basic Engineering* **82**, 34-45.

[23] KALMAN, R.E. & BUCY, R.S. (1961). "New Results in Linear Filtering and Prediction Theory", *J. of Basic Engineering* **83**, 95-108.

[24] KALMAN, R.E., FALB, P.L. & ARBIB, M.A. (1969). *Topics in Mathematical System Theory*, McGraw-Hill.

[25] LEE, A.T. (1980). "A Direct Approach to Identify the Noise Covariances of Kalman Filtering", *IEEE Trans. Automat. Contr.* **AC-25**, 841-842.

[26] LJUNG, G.M. & BOX, G.E.P. (1978). "On a measure of lack of fit in time series models", *Biometrika* **65**, 297-303.

[27] MEINHOLD, R.J. & SINGPURWALLA, N.D. (1983). "Understanding the Kalman Filter", *The Amer. Statist.* **37**, 123-137.

[28] MEHRA, R.K. (1970). "On the Identification of Variances and Adaptive Kalman Filtering", *IEEE Trans. Automat. Contr.* **AC-15**, 175-184.

[29] MEHRA, R.K. (1972). "Approaches to Adaptive Filtering", *IEEE Trans. Automat. Contr.* **AC-17**, 693-698.

[30] MEHTA, B.M., AHLERT, R.C. & YU, S.L. (1975). "Stochastic Variation of Water Quality of the Passaic River", *Water Resources Research* **11**, 300-8.

[31] MIGON, H.S. & HARRISON, P.J. (1985). "An Application of Non-linear Bayesian Forecasting to Television Advertising", in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.), North-Holland, Amsterdam, and Valencia University Press.

[32] MILHOJ, A. (1981). "A test of fit in time series models", *Biometrika* **68**, 177-87.

[33] PIERCE, D.A. (1977). "Relationships - and the Lack Thereof - Between Economic Time Series, with Special Reference to Money and Interest Rates", *Journal of the American Statistical Association* **72**, 11-26.

[34] POSKITT, D.S. & TREMAYNE, A.R. (1981). "An Approach to Testing Linear Time Series models", *The Annals of Statistics* **9**, 974-86.

[35] StatSci Europe, Statistical Sciences U.K. Ltd. (1992). *S-plus Programmer's Manual.*

[36] WEST, M. & HARRISON, P.J. (1989). *Bayesian Forecasting and Dynamic Models*, Springer-Verlag New York Inc.

[37] WEST, M., HARRISON, P.J. & MIGON, H.S. (1985). "Dynamic Generalised Linear Models and Bayesian Forecasting", *Journal of the American Statistical Society* **80**, 73-97.

[38] WINTERS, P.R. (1960). "Forecasting sales by exponentially weighted moving averages", *Management Science* **6**, 324-342.

[39] YOUNG, P.C. (1974). "Recursive Approaches to Time Series Analysis", *Bull. Inst. Maths. and Applicns.* **10**, 209-224.

Appendix A

Time series used in text

Series 1: Total bakery sales of white and brown bread (daily data, not including Sundays, total of 44 weeks - read across); missing values denoted by 0. Christmas week is indicated by underlining.

Mon	Tues	Weds	Thur	Fri	Sat	Mon	Tues	Weds	Thur	Fri	Sat
1496	1494	2000	2788	3688	2491	1634	1601	2275	3054	4077	2409
0	2167	2314	2962	3797	2979	1665	1601	1947	3007	3513	2441
1525	1635	2241	2819	3484	2065	1473	1669	2183	3052	3931	2236
0	1994	1861	2548	3529	2266	1522	1453	2055	2752	3687	2329
1495	1506	1937	2719	3594	2241	1215	1444	1827	2723	3607	2358
1509	1429	2018	2796	3662	2323	1493	1618	1825	2580	3493	2159
1625	1390	2080	2764	3754	2141	1378	1566	1887	2729	3153	1836
1437	1370	1877	2577	3220	2137	1516	1468	1703	2408	3092	1642
1569	1474	1764	2470	3057	1216	1479	1409	1765	2626	3120	1593
1601	1504	1910	2806	3062	1962	1601	2190	1982	2634	3675	2230
1625	1522	1975	2945	3718	2178	1658	1536	2011	2890	3671	2562
1627	1635	1990	2859	3755	1992	1619	1539	2002	3018	3723	2060
1793	1790	2009	3050	3773	2661	1862	1591	2002	2977	3833	2409
1831	1718	2169	2976	3784	2657	1835	1762	2237	2973	3677	2649
1824	1744	1986	2847	3531	2276	1858	1593	2075	2850	3579	2303
1811	1668	2159	2983	3748	2550	1800	1663	2111	2998	3724	2716
1876	1556	2087	2844	3720	2548	1739	1598	2116	2743	3564	2589
1705	1723	2143	2734	3584	2646	2157	2205	3430	4626	3366	2365
0	0	1289	2085	2869	1922	0	2138	2106	2894	3670	2440
1877	1679	2149	3068	3771	2589	1941	1753	2151	3037	3807	2452
1890	1849	2172	2962	3955	3001	1976	1789	2195	3116	3911	2673
1861	1704	1988	2983	3774	2565	1923	1743	2265	3122	3889	2662

Series 2: Advertising awareness data, X_t and $\mu_t = Y_t/n_t$.

X_t : TVR units (weekly, by row)														
0.05	0.00	0.20	7.80	6.10	5.15	1.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.50	4.60	3.70	1.45	1.20	2.00	3.40	4.40	3.80	3.90	5.00	0.10	0.60	3.85	3.50
3.15	3.30	0.35	0.00	2.80	2.90	3.40	2.20	0.50	0.00	0.00	0.10	0.85	4.65	5.10
5.50	2.30	4.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.0	10.0	9.00	8.00	7.00	6.00
5.00	4.00	3.00	2.00	1.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.0
10.0	9.00	8.00	7.00	6.00	5.00	4.00	4.00	5.00	6.00	7.00	8.00	9.00	10.0	11.0
12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	20.0	19.0	18.0	17.0	16.0	15.0
14.0	13.0	12.0	11.0	11.0	12.0	2.90	4.47	2.24	6.71	3.22	7.29	6.66	2.48	3.64
8.70	7.11	2.30	6.40	3.20	7.11	3.57	7.93	5.13	6.40	3.31	2.68	5.39	7.22	4.40
6.69	8.69	6.32	5.99	6.94	6.19	7.59	3.59	8.44	8.61	8.08	9.32	9.13	8.39	9.58

μ_t : Awareness response proportion, Y_t/n_t														
0.40	0.41	0.31	0.40	0.45	0.44	0.39	0.50	0.32	0.42	0.33	0.24	0.25	0.32	0.28
0.25	0.36	0.38	0.36	0.29	0.43	0.34	0.42	0.50	0.43	0.43	0.52	0.45	0.30	0.55
0.33	0.32	0.39	0.32	0.30	0.44	0.27	0.44	0.30	0.32	0.30	0.00	0.00	0.00	0.33
0.48	0.40	0.44	0.40	0.34	0.37	0.37	0.23	0.30	0.21	0.23	0.22	0.25	0.23	0.14
0.21	0.16	0.19	0.07	0.26	0.16	0.21	0.07	0.22	0.10	0.15	0.15	0.22	0.11	0.14
0.04	0.19	0.19	0.29	0.36	0.40	0.28	0.43	0.57	0.58	0.59	0.67	0.50	0.63	0.66
0.61	0.48	0.65	0.30	0.50	0.41	0.51	0.36	0.44	0.47	0.39	0.48	0.40	0.50	0.61
0.58	0.39	0.68	0.47	0.70	0.52	0.45	0.59	0.57	0.49	0.42	0.51	0.59	0.63	0.68
0.61	0.70	0.63	0.59	0.67	0.66	0.81	0.75	0.51	0.66	0.68	0.55	0.74	0.56	0.65
0.64	0.66	0.57	0.56	0.62	0.72	0.00	0.00	0.00	0.00	0.00	0.54	0.55	0.53	0.52
0.54	0.55	0.53	0.54	0.52	0.54	0.52	0.54	0.54	0.55	0.53	0.51	0.52	0.53	0.53
0.54	0.56	0.56	0.56	0.56	0.56	0.57	0.55	0.56	0.58	0.58	0.59	0.60	0.60	0.61

Series 3: Simulated 1st-order TSDLM (steady model) $M_0 = \{1, 1, 1, 0.5\}$, with $m_0 = 10$; first 200 points of series, overall length 1000.

9.97	9.15	9.00	8.04	8.91	6.67	6.95	7.20	7.48	8.94	10.49	11.14	7.38	12.03
10.52	10.45	12.88	10.94	10.92	11.71	10.97	11.02	9.82	9.80	14.19	11.37	12.00	11.67
10.38	8.89	8.80	9.30	9.46	7.97	11.58	10.80	12.98	9.75	10.37	11.76	13.62	10.67
10.63	12.60	13.81	13.77	10.13	11.62	9.82	12.27	9.66	8.13	7.39	7.14	7.84	8.21
8.44	8.86	6.63	5.04	5.97	5.32	5.21	4.66	2.42	3.19	4.43	3.94	5.93	6.37
5.10	6.15	4.52	3.44	5.51	7.32	5.99	6.60	2.48	4.82	5.26	2.30	5.04	6.22
6.37	8.54	5.77	7.33	5.74	8.20	6.98	6.90	5.56	7.22	8.13	6.48	7.08	7.33
7.93	7.38	5.61	6.40	9.45	7.36	8.25	8.81	9.09	8.27	8.14	7.10	6.84	6.64
5.73	4.68	2.64	4.02	5.60	5.36	6.06	6.40	6.08	2.39	5.38	4.63	6.09	5.90
3.77	3.92	1.82	3.40	3.64	2.14	3.40	2.92	3.66	4.60	5.03	5.18	5.21	3.90
6.10	6.47	7.98	7.38	7.58	4.80	5.73	6.02	7.31	6.02	7.47	4.59	5.16	3.95
5.94	5.76	7.45	5.44	5.01	7.10	8.43	5.44	7.63	7.43	6.86	8.25	9.80	8.95
8.72	11.09	10.89	10.45	11.06	9.60	10.97	10.42	9.88	11.99	12.89	13.92	12.17	14.51
12.36	13.89	15.65	13.67	12.92	12.05	11.83	12.00	12.56	13.45	13.97	12.13	12.90	13.20
13.40	13.23	12.83	10.87										

Series 4: Concentration readings from a chemical process (as in Box and Jenkins, Series A, p.525).

17.0	16.6	16.3	16.1	17.1	16.9	16.8	17.4	17.1	17.0	16.7	17.4	17.2	17.4	17.4
17.0	17.3	17.2	17.4	16.8	17.1	17.4	17.4	17.5	17.4	17.6	17.4	17.3	17.0	17.8
17.5	18.1	17.5	17.4	17.4	17.1	17.6	17.7	17.4	17.8	17.6	17.5	16.5	17.8	17.3
17.3	17.1	17.4	16.9	17.3	17.6	16.9	16.7	16.8	16.8	17.2	16.8	17.6	17.2	16.6
17.1	16.9	16.6	18.0	17.2	17.3	17.0	16.9	17.3	16.8	17.3	17.4	17.7	16.8	16.9
17.0	16.9	17.0	16.6	16.7	16.8	16.7	16.4	16.5	16.4	16.6	16.5	16.7	16.4	16.4
16.2	16.4	16.3	16.4	17.0	16.9	17.1	17.1	16.7	16.9	16.5	17.2	16.4	17.0	17.0
16.7	16.2	16.6	16.9	16.5	16.6	16.6	17.0	17.1	17.1	16.7	16.8	16.3	16.6	16.8
16.9	17.1	16.8	17.0	17.2	17.3	17.2	17.3	17.2	17.2	17.5	16.9	16.9	16.9	17.0
16.5	16.7	16.8	16.7	16.7	16.6	16.5	17.0	16.7	16.7	16.9	17.4	17.1	17.0	16.8
17.2	17.2	17.4	17.2	16.9	16.8	17.0	17.4	17.2	17.2	17.1	17.1	17.1	17.4	17.2
16.9	16.9	17.0	16.7	16.9	17.3	17.8	17.8	17.6	17.5	17.0	16.9	17.1	17.2	17.4
17.5	17.9	17.0	17.0	17.0	17.2	17.3	17.4	17.4	17.0	18.0	18.2	17.6	17.8	17.7
17.2	17.4													

Series 5: Simulated 2-dimensional constant TSDLM, from
$M_0 = \left\{(1,1)^T, \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, 1, \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}\right\}$, with $\mathbf{m}_0 = (1,1)^T$.

1.55	1.37	2.54	1.94	0.32	-0.20	2.18	0.95	2.18	2.94	2.32	1.53	1.89	0.94
0.87	-3.10	-1.10	1.54	0.76	3.33	-1.08	-0.47	1.51	3.03	3.97	3.75	4.17	2.59
1.16	0.38	1.59	1.46	1.50	0.41	0.58	2.89	1.44	4.44	1.17	3.76	2.96	5.98
5.94	5.57	6.21	5.33	6.20	7.69	5.56	10.61	8.32	4.09	7.85	10.16	10.03	9.38
10.30	10.11	9.32	9.61	10.29	10.14	9.42	9.67	8.67	6.31	7.74	6.33	8.16	6.79
6.42	6.52	5.92	8.02	9.37	11.07	8.67	8.86	6.86	9.48	8.88	8.26	8.08	9.77
8.73	5.66	9.58	5.30	7.91	6.61	9.91	9.20	7.23	6.77	6.72	9.96	9.00	10.55
13.19	11.85	11.60	10.83	7.26	8.20	9.44	8.51	5.29	5.43	8.26	9.66	7.09	-1.28
1.88	1.13	3.61	2.96	3.14	2.49	4.91	4.65	3.59	3.49	4.52	1.71	4.53	1.40
4.63	4.13	5.66	7.50	9.96	7.76	4.81	6.61	6.99	6.18	4.43	6.79	7.35	6.12
7.31	8.10	7.55	7.84	7.95	6.78	5.96	8.85	3.89	7.45	7.11	7.36	7.06	5.53
4.56	9.44	7.11	5.93	5.94	6.02	6.87	4.88	5.74	5.05	7.07	6.42	3.46	4.05
3.21	4.55	4.12	4.39	5.98	6.26	6.48	10.01	9.24	8.22	6.23	8.12	7.20	11.74
6.22	5.95	8.72	9.31	8.06	6.57	8.28	8.52	8.97	3.71	7.80	6.49	6.15	6.77
4.06	5.30	5.32	5.25										